THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

In search of non-photosynthetic Cyanobacteria

Rochelle Melissa Soo

BCA, BBMedSc, Victoria University of Wellington

MSc (Hons I), University of Waikato

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2015

School of Chemistry and Molecular Biosciences

Abstract

One of the major evolutionary events that occurred on our planet is the establishment of organisms capable of performing oxygenic photosynthesis. This changed the earth's atmosphere from an anoxic to an oxic environment which likely contributed to the development of more complex organisms. Cyanobacteria are the only known prokaryotes capable of performing oxygenic photosynthesis and until recently, it was believed that all Cyanobacteria carry out this process. With the advent of culture-independent molecular techniques, a number of basal lineages of Cyanobacteria have been detected and classified as 4C0d-2 and ML635J-21. Many representatives of these lineages have been found in aphotic environments, raising the possibility that they are non-photosynthetic. The main aim of this thesis was to use metagenomics to obtain genomes belonging to the cyanobacterial basal lineage 4C0d-2, determine if they contain the photosynthetic apparatus required for oxygenic photosynthesis and whether they should be classified as Cyanobacteria or as a separate phylum.

In Chapter Two, amplicon pyrosequencing was used to identify environments that contain members of the basal lineage, 4C0d-2. Positive habitats were shotgun sequenced and six genomes were extracted from 4C0d-2 populations using differential coverage binning; three from koala faeces, one from human faeces, one from a lab-scale (EBPR), and one from a full-scale (UASB). No photosynthetic genes were identified in any of the 4C0d-2 genomes suggesting that this lineage is indeed non-photosynthetic. Genome-based phylogenetic trees confirmed that 4C0d-2 shares a common ancestor with photosynthetic cyanobacteria. An independent study by Di Rienzi et al (2013) concluded that 4C0d-2 is a sister phylum of the Cyanobacteria for which they proposed the name Melainabacteria. Based on the robust phylogenetic association and a number of inferred common traits between the two groups, I proposed that the Melainabacteria should be reclassified as a class within the Cyanobacteria, comprising four orders represented by genome sequences; Gastranaerophilales, Obscuribacterales, Caenarcaniphilales and Vampirovibrionales.

During analysis of Melainabacteria 16S rRNA genes, the sequence of a possible cultured representative, *Vampirovibrio chlorellavorus,* was discovered (after which I named one of the orders above). In the 1970's, *V. chlorellavorus* was observed preying upon the microalga, *Chlorella vulgaris* and was initially classified as a *Bdellovibrio* (member of the Deltaproteobacteria) based on its predatory nature and cell shape. In Chapter Three, the DNA from 36 year-old lyophilised cells of *V. chlorellavorus* and *C. vulgaris* were extracted and the genome of *V. chlorellavorus* was shotgun

sequenced and assembled into a near-complete draft genome. Genome-based trees confirmed that *V. chlorellavorus* is a member of the Melainabacteria and not the Deltaproteobacteria, expanding the number of known phyla containing predatory bacteria to five. The molecular machinery used by *V. chlorellavorus* for predation was inferred from the annotated genome.

Representatives from the order Obscuribacterales were detected in a sample collected from intact permafrost (palsa) as part of another study. In Chapter Four, two genomes from this order were extracted from ultra-deep metagenomic sequencing and metabolic reconstructions were created to provide an in-depth insight into their functionality. Metatranscriptomic analysis of one palsa sample was performed to determine which genes are actively expressed by the permafrost Obscuribacterales.

The findings presented in this thesis provide a useful basis for understanding the newly discovered non-photosynthetic cyanobacterial lineage, the Melainabacteria, and increases the number of known phyla containing predatory bacteria. For example, the Melainabacteria genomes may be used to understand the evolutionary origins of oxygenic photosynthesis, design probes for visualisation or identify potential media for culturing.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Published peer-reviewed papers

Soo, R.M., Skennerton, C.T., Sekiguchi, Y., Imelfort, M., Paech, S.J., Dennis P,G., Steen, J.A., Parks, D.H., Tyson, G.W., and Hugenholtz, P. 2014. An Expanded Genomic Representation of the Phylum Cyanobacteria. *Genome Biology and Evolution*. *6 (5): 1031-1045.*

Soo, R.M., Woodcroft, B.J., Parks, D.H., Tyson, G.W., and Hugenholtz, P. 2015. Back from the dead; the curious tale of the predatory Cyanobacterium *Vampirovibrio chlorellavorus*. *PeerJ. 3:e968.*

Publications included in this thesis

Soo, R.M., Skennerton, C.T., Sekiguchi, Y., Imelfort, M., Paech, S.J., Dennis P,G., Steen, J.A., Parks, D.H., Tyson, G.W., and Hugenholtz, P. 2014. An Expanded Genomic Representation of the Phylum Cyanobacteria. *Genome Biology and Evolution*. *6 (5): 1031-1045* - incorporated as Chapter 2.

Contributor	**Statement of contribution**
Soo, R.M. (Candidate)	Designed experiments (30%) Performed experiment (28%) Analysed data (70%) Wrote and edited paper (60%)
Skennerton, C.T.	Designed experiments (10%) Performed experiment (35%) Analysed data (20%) Wrote and edited paper (2%)
Sekiguchi, Y.	Designed experiments (10%) Performed experiment (35%) Analysed data (20%) Wrote and edited paper (1%)
Imelfort, M.	Analysed data (2%)
Paech, S.J.	Analysed data (2%)

Dennis, P.G.	Performed experiment (2%)
Steen, J.A.	Analysed data (1%)
Parks, D.H.	Analysed data (5%) Wrote and edited the paper (2%)
Tyson, G.W.	Designed experiments (20%) Wrote and edited the paper (15%)
Hugenholtz, P.	Designed experiments (30%) Wrote and edited paper (20%)

Soo, R.M., Woodcroft, B.J., Parks, D.H., Tyson, G.W., and Hugenholtz, P. 2015. Back from the dead; the curious tale of the predatory Cyanobacterium *Vampirovibrio chlorellavorus*. *PeerJ. 3:e968* - incorporated as Chapter 3.

Contributor	**Statement of contribution**
Soo, R.M. (Candidate)	Designed experiments (60%) Analysed data (75%) Wrote and edited paper (60%)
Woodcroft, B.J.	Designed experiments (5%) Analysed data (10%) Wrote and edited paper (1%)
Parks, D.H.	Designed experiments (5%) Analysed data (15%) Wrote and edited paper (2%)
Tyson, G.W.	Designed experiments (10%) Wrote and edited paper (2%)
Hugenholtz, P.	Designed experiments (20%) Wrote and edited paper (35%)

Contributions by others to the thesis

Chapter 4 – Population genomics and transcriptomics of two Obscuribacterales populations recovered from palsa in Stordalen Mire, northern Sweden

Contributor	Statement of contribution
Soo, R.M. (Candidate)	Analysed data (75%) Wrote and edited chapter (93%)
Woodcroft, B.J.	Designed experiments (10%) Analysed data (10%)
Singleton, C.	Analysed data (15%)
Hugenholtz, P.	Wrote and edited chapter (5%)
Tyson, G.W.	Designed experiments (90%) Wrote and edited chapter (2%)

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor, Phil Hugenholtz, who took me under his wing as a PhD student after a Skype conversation. He has taught me to think critically and pushed me to think outside the box. Thank you for all the countless hours that were spent discussing different parts of my thesis and going through the manuscripts. I think at times it has been a bit of a roller coaster ride but we finally made it! I would also like to thank my co-supervisor, Gene Tyson, for his support, advice and guidance.

Thank you to my thesis committee panel members, Scott Beatson, Al McKinnon and Bernie Degnan for their helpful advice and questions during my PhD milestones. I thanks my fellow lab mates at ACE, especially Nancy Lachner, Caitlin Singleton, Dyana Rahman, Serene Lowe, Josh Daly, Fauzi Haroon, Connor Skennerton, Steve Robbins and Inka Vanwonterghem. It was great to have other PhD students to discuss the lows and highs of writing a thesis and know that we were all going through the same thing. A big thank you to Donovan Parks and Ben Woodcroft for their advice and guidance. Thanks to my friends Ana Cano Gomez and Katie Glover for their support.

I'd like to thank my family, especially my husband Brett, who was willing to move to Brisbane with me so I could continue my study. He has been my rock during all the ups and downs. Last but not least, I would like to thank my family: my parents and my brother, who have supported me throughout writing this thesis. Also thank you to all my extended family for the enjoyable small breaks during the last 3 and a half years.

I'm also grateful to the ARC for providing me with an APA scholarship that financially supported me during my PhD and for the grant that was awarded to Phil to carry out the research.

Keywords

Cyanobacteria, Melainabacteria, comparative genomics, photosynthesis, metabolism, phylogenetics

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060408, Genomics, 50%
ANZSRC code: 060309, Phylogeny and Comparative Analysis, 20%
ANZSRC code: 060104, Cell Metabolism, 30%

Fields of Research (FoR) Classification

FoR code: 0605, Microbiology, 70%
FoR code: 0601, Biochemistry and cell biology, 30%

Table of Contents

List of Figures

List of Tables

List of Abbreviations used in the thesis

ANI	Average Nucleotide Identity
ATCC	American Type Culture Collection
EBPR	Enhanced Biological Phosphorus Removal
EMP	Embden-Meyerhof-Parnas pathway
IMG/ER	Integrated Microbial Genomes/Expert Review
KEGG	Kyoto Encyclopedia of Genes and Genomes
NCIMB	National Collection of Industrial, Marine and Food Bacteria
PCR	Polymerase Chain Reaction
PS	Photosystems
RC	Reaction Centres
TCA	Tricarboxylic acid cycle
TFP	Type IV pili
T4SS	Type IV secretion system
UASB	Upflow Anaerobic Sludge Blanket

Chapter 1: Literature Review

1.1 Cyanobacteria

As one of the most abundant and diverse groups of microorganisms, members of the phylum Cyanobacteria play a major role in the production of oxygen through the process of oxygenic photosynthesis. These organisms changed the Earth's atmosphere from a reducing to an oxidising one, which led to the creation of more complex organisms [1]. In addition to performing photosynthesis, Cyanobacteria also play a key role in carbon and nitrogen cycles [2]. They are a major contributor towards CO_2 sequestration with the products from photosynthesis, ATP and NADPH fixing atmospheric CO_2 via the Calvin cycle into carbon skeletons for the synthesis of starch and sucrose [3]. Many Cyanobacteria are capable of fixing nitrogen using either heterocysts (microoxic cells that provide an environment in oxic environments) or akinetes (reproductive spores), which protect the highly oxygen-sensitive nitrogenase [4]. Nitrogen-fixing Cyanobacteria that lack heterocysts or akinetes have the ability to alternate their carbon metabolism between oxygenic photosynthesis and an anoxygenic form using sulphide as an electron donor, instead of water to allow nitrogen fixation [5]. In the case of cyanobacterium *Candidatus* Atelocyanobacterium thalass [6], an absence of photosystem II (the protein complex required for oxygenic photosynthesis) allows the bacteria to express nitrogenase genes during photosynthesis [7].

Cyanobacteria can be found in almost every aquatic and terrestrial habitat including freshwater [8], marine [9], desert crusts [10] and endolithic borings in rocks [11]. They are morphologically diverse and can range from single-cells to filaments [12] and they have an unusual Gram-negative cell wall which is thicker (10 to > 700 nm) in comparison with other Gram-negative bacteria (2 to 6 nm). Although they do not possess flagella, some are able to utilise a peculiar type of motility, called gliding [13].

Genome sizes can vary by almost an order of magnitude, ranging from a minimum of 1.44 Mbp for the streamlined marine cyanobacterium *Candidatus* Atelocyanobacterium thalassa [7] to *Scytonema hofmanni* PCC 7110, with 11.96 Mbp [14]. The number of genes found in Cyanobacteria ranges from 1,241 to 12,356 from *S. hofmanni* PCC 7110, which is the most gene-rich prokaryote currently known [14]. Furthermore, most contain circular genomes with additional plasmids [4]. Cyanobacteria can also differ in ploidy (genome copy number) with a real-time PCR study showing that *Synechococcus* PCC 7942 and WH7803 have 3 to 4 genome copies per cell, whereas

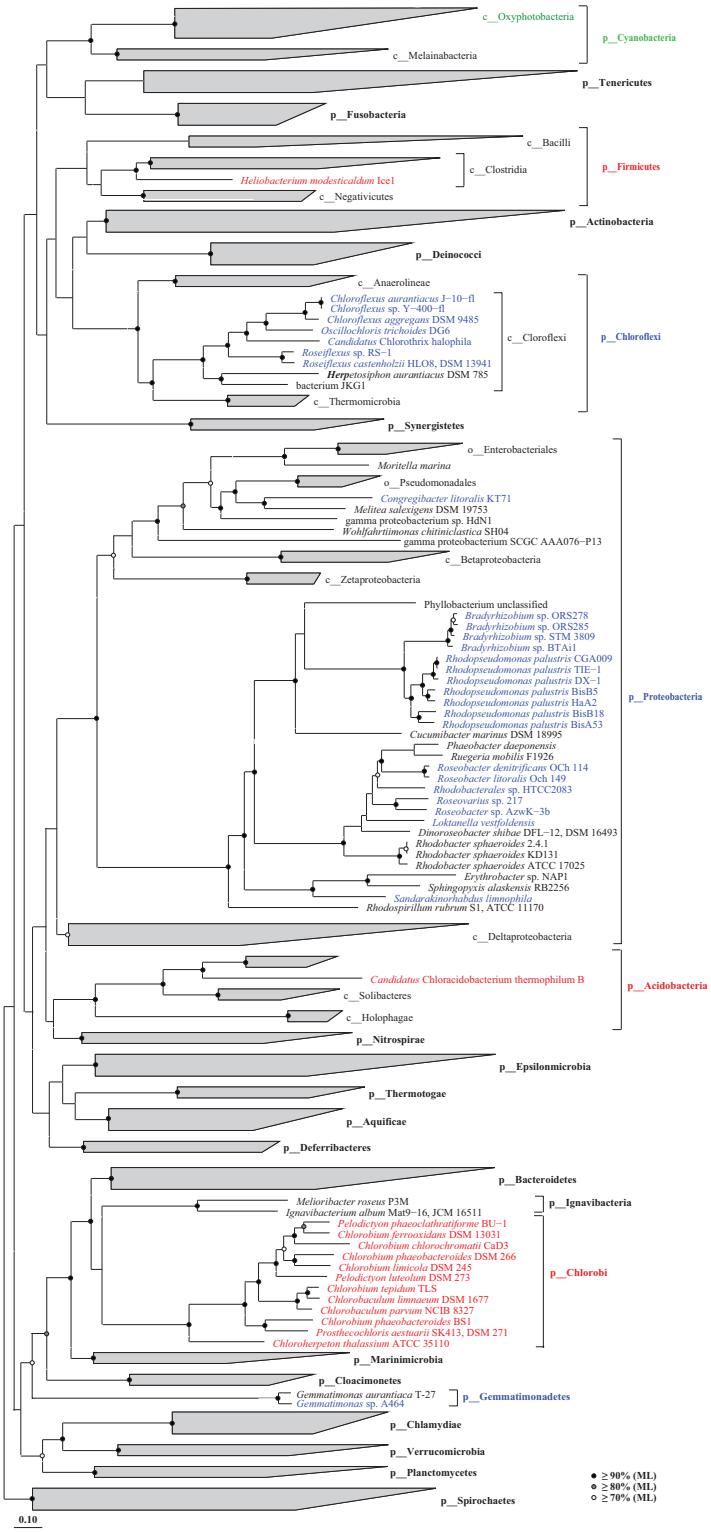
Synechocystis PCC 6803 contains 218 genome copies in exponential phase and 42 genome copies in linear and stationary growth phase [15].

1.2 Photosynthesis

Photosynthesis is the conversion of solar energy by plants, most algae and certain bacteria for the synthesis of complex organic molecules needed to power life. There are two types of photosynthesis, oxygenic and anoxygenic [16]. In the oxygenic process, water is used as the electron donor with oxygen as a byproduct, whereas in anoxygenic photosynthesis, organic or sulphur compounds, nitrite, arsenite, and molecular hydrogen can be used as electron donors, however oxygen is not produced [17].

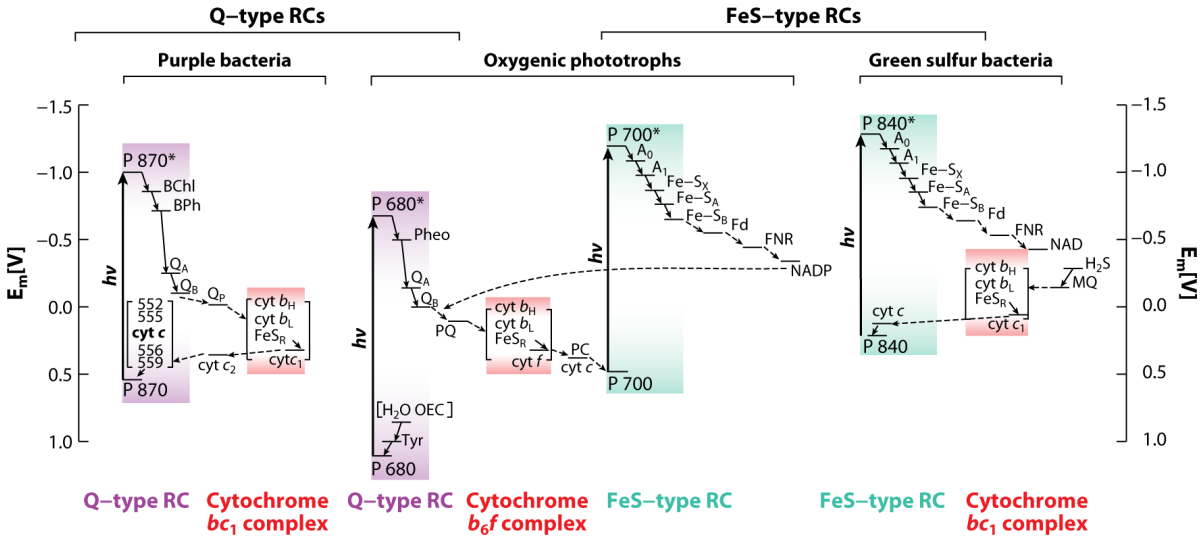
Within the bacteria, photosynthetic organisms are found in the following phyla: Cyanobacteria, Proteobacteria, Chloroflexi, Firmicutes, Chlorobi, Acidobacteria and the recently added Gemmatimonadetes [18] (**Figure 1.1**). The core of the photosynthetic apparatus called the photosystems (PS), contain the reaction centres (RC) in which the chlorophylls or bacteriochlorophylls (the principal antennae pigments), as well as other cofactors such as quinone or iron sulphur centres are located and charge separation occurs. Chlorophyll, required for oxygenic photosynthesis, and bacteriochlorophyll, required for anoxygenic photosynthesis are used to absorb light, which is then channelled to RCs where they perform photochemical charge separation and send electrons through the electron transport chain. In addition, all photosynthetic organisms contain different chlorophylls that also differ in structure, allowing them to absorb light at different frequencies [19]. The RCs initiate light-driven electron transport and are classified into two categories, RC1 and RC2, depending on the electron acceptor [17]. RC1 has an iron-sulfur (FeS) type electron acceptor and is found in photosynthetic Chlorobi, Firmicutes, Acidobacteria and PSI of Cyanobacteria. RC1 contains chlorophyll A molecules as a fundamental part of the charge separation and electron transfer to ferredoxin. RC2 has a pheophytin-quinone type (Q-type) electron acceptor and is found in photosynthetic Proteobacteria, Chloroflexi, Gemmatimonadetes and PSII of Cyanobacteria (**Figure 1.2)**. Cyanobacteria are the only bacteria to possess both RC1 and RC2 and can therefore uniquely perform oxygenic photosynthesis, whereas the other photosynthetic bacteria contain RC1 or RC2 and perform anoxygenic photosynthesis [17].

Figure 1.1. Maximum likelihood phylogenetic tree of phyla with photosynthetic and non-photosynthetic representatives



Species in red encode genes for RC1, species in blue encode genes for RC2 and species in green encode genes for RC1 and RC2. p__ represents phylum, c__ represents class and o__ represents order. Black circles in the tree represents nodes with >90% bootstrap support, grey circles represent nodes with >80% bootstrap support and white circles in the tree represent nodes with >70% bootstrap support.

Figure 1.2. Overview of reaction centre types in Proteobacteria (purple bacteria), Cyanobacteria (oxygenic phototrophs) and Chlorobi (green sulphur bacteria)



The Proteobacteria contain Q-type RC (RC2), the Chlorobi contain Fe-S type RCs (RC1) and Cyanobacteria contain both a Q-type (RC2) and a FeS-type (RC1). Image from Hohmann-Marriott and Blankenship, 2011.

1.2.1 The origin and evolution of photosynthesis

The origin of photosynthesis has long been debated and in particular the question of the earliest ancestral photosynthetic organism [20]. Isotopic evidence, fossils, geochemical analysis and biochemical markers have been used to try to predict the origin of photosynthesis [21]. Isotopic evidence for carbon fixation date back to 3.8 billion years ago (or Gig-annum, Ga) [22], however, this evidence has been questioned by many as Fischer-Tropsch synthesis of organic compounds at hydrothermal systems can produce molecules with a similar C isotope ratio [23]. The earliest Cyanobacteria-like microfossil shows that these organisms were likely present 3.2 to 3.5 Ga in stromatolites, layered structures consisting of mat-forming organisms and sediment. Notably, the stromatolite findings remain controversial because it is impossible to determine cell physiology from microfossils [24]. The oldest cyanobacterial fossils generally accepted as Cyanobacteria, date back to ~1.9 Ga [25]. Other fossils that have been identified include those described by Schopf who identified fossils preserved in Early Archean Apex Basalt but these have been widely dismissed [26]; filaments preserved in sulphide described by Rasmussen and others and organic walled fossils discovered in the Moodies by Javaux *et al.* [27]. Chemical biomarkers (2-methylhopanoids) found in ancient rock suggest that cyanobacteria-like organisms existed before 2.5 Ga, however it is not possible to determine if these are from Cyanobacteria [28, 29] and contamination is a huge issue in

this field [30]. The most prominent evidence for phototrophy comes from Tice and Lowe who showed evidence of photosynthetic carbon fixation by filamentous microbial mats found in 3.4 Ga sedimentary rocks in anoxic environments. Their geochemical analysis identified hydrogen as the primary electron source, suggesting that early photosynthesis was carried out by anoxygenic photosynthetic organisms [31].

Three lines of evidence have shown that the atmosphere was increasing in oxygen by 2.4 Ga: the nitrogen-oxygen redox cycle establishes oxygen increasing by 2.7 Ga; the chromium signatures suggest the presence of oxygen by 2.8 Ga and the sulphur fractionation data suggests oxygen accumulation by 2.5 to 2.45 Ga. It has been predicted that the accumulation of oxygen occurred in two phases. The first phase occurred between 2.4 and 2.0 Ga during the "Great Oxidation Event" (GOE), with a gradual increase in atmospheric oxygen to 1-2%. Oxygen layers may have been stable before rising to the 20% levels found in today's atmosphere. The second rise occurred with the emergence of photosynthetic eukaryotes [17]. However, indications of oxygen being produced before this time have been shown by banded iron formations (BIFs), which suggest that the oxidation of Fe^{2+} to Fe^{3+} may have occurred 3.5 Ga, with biogenic oxygen being liberated for ~1 billion years before it started to accumulate in the atmosphere. These data suggest that Cyanobacteria have been performing oxygenic photosynthesis since 3.5 Ga. An alternative to this theory is that BIFs may have been produced by anoxygenic photosynthetic organisms that utilise ferrous iron as an electron donor [16], which would suggest that oxygenic Cyanobacteria may only be 2.5 Ga. Most of the geological evidence for the rise of oxygen remains controversial. However, the S isotopes and redox sensitive detritral grains [32] are thought to be the most reliable evidence, showing that the rise of oxygen occurred ~2.4-2.3 Ga.

The presence of RC1 and RC2 in Cyanobacteria has led many to question how both RCs arose in this organism (**Figure 1.2**). Two main models have been put forward for the evolution of the RCs, the selective loss model and the fusion model. In the selective loss model, it is suggested that both RCs were present in a single organism and with the exception of the Cyanobacteria, photosynthetic organisms lost either the FeS or Q-type RC. In the fusion model, it has been proposed that both RCs developed in different organisms and a bacterium that gave rise to the cyanobacterial line already contained one RC but was provided with an additional RC through lateral gene transfer [17]. To date, the most widely accepted model is the fusion model due to the greater simplicity of the subunit composition of the RCs in anoxygenic photosynthetic organisms [33]. However, Sousa and colleagues argue against the fusion model as their study suggests the presence of two serially linked photosystems at the origin of water-splitting photosynthesis [34]. Other studies have suggested a

duplication event [35, 36]. Whether RC1 or RC2 organisms were the first to photosynthesise is still under debate, although a number of studies have suggested that RC1-containing Chlorobi were the original photosynthetic organisms [37, 38]. Perhaps declaring whether RC1 or RC2 was first may be too simplistic as arguments against both models can be made.

Another question that has been hotly debated is how did photosynthesis evolve in diverse lineages? Photosynthetic organisms represent mosaics of genes with different evolutionary histories so identifying genes that can be used to reliably infer the evolution of photosynthesis is challenging [39]. Despite this issue, studies of chlorophyll and bacteriochlorophyll, as well as RCs have been used to try to answer this question. The evolution of chlorophyll has been used as an indicator for the evolution of photosynthesis as a whole. The most popular school of thought about how photosynthesis arose in multiple divergent lineages is the Granick hypothesis, which states that the evolution of the chlorophyll biosynthetic pathway followed the sequential inventions of new enzymes to generate more stable products [40]. However the topic is still under debate with Xiong and colleagues providing evidence through phylogeny for the bacteriochlorophyll/chlorophyll genes that chlorophyll-a biosynthesis evolved from a more complex bacteriochlorophyll biosynthetic pathway with Proteobacteria identified as the earliest emerging photosynthetic lineage [37]. On the other hand, these results remain controversial because the bacteriochlorophyll/chlorophyll proteins provide different results depending on the evolutionary model and alignments used and the proteins are small, providing little evolutionary information. In addition, there is also a large amount of reticulate evolution and paralogous evolution in the pigment pathway, making it difficult to reconstruct the right taxonomy history. Over 100 genes are needed for the synthesis and regulation of the photosynthetic apparatus, however the pattern of the gene order differs amongst photosynthetic taxa. In Proteobacteria and Firmicutes, most of the photosynthetic genes are clustered in large contiguous photosynthetic gene clusters, whereas in Chloroflexi, Chlorobi and Cyanobacteria, small clusters of two to four genes are conserved in linkage. In addition, the gene organisation in the Proteobacteria suggests lateral gene transfer. It is thought that the divergent clustering patterns of photosynthetic genes in Cyanobacteria have emerged as a result of progressive operon splitting, which may lead to different gene recombinations [41].

Non-photosynthetic markers have also been used to predict the earliest photosynthetic organism. One study used the conservative 16S rRNA gene to create phylogenetic trees and identified Chloroflexi as the earliest photosynthetic lineage, followed by Heliobacteria, Chlorobi, Cyanobacteria and Proteobacteria [19]. In another study, Heliobacteria is identified as the earliest branching of the photosynthetic bacteria based on 16S rRNA [36], perhaps highlighting the limited

resolution of inter-phylum branching orders in the bacterial domain [42]. Gupta and colleagues used heat shock proteins to construct phylogenetic trees of photosynthetic organisms, and the presence of shared insertions and deletions (indels) to infer the order of evolutionary events. They concluded that Heliobacteria were the most ancient, followed by Chloroflexi, Cyanobacteria, Chlorobi and Proteobacteria. A study by Mulkidjanian and colleagues analysed 15 cyanobacterial genomes and their photosynthetic genes, comparing them to other photosynthetic organisms and concluded that Cyanobacteria were the most ancestral phototroph due to their enlarged photosynthetic core gene set. They suggested that a group termed 'procyanobacteria', an ancestor of the present day Cyanobacteria, was the most ancient phototroph and they spread their photosynthetic genes to other phyla by horizontal gene transfer (HGT) [43]. Based on these disparate lines of evidence for the origin of photosynthesis, the jury is clearly still out on the identity of the earliest photosynthetic ancestor.

1.3 Classification of Cyanobacteria

Cyanobacteria were first defined in the early 19th century as a class or division of algae due to their photosynthetic capabilities. It was not until 1962 when the differences between eukaryotes and prokaryotes were defined that the Cyanobacteria were reclassified as prokaryotes [44]. Cyanobacteria have historically been classified according to the Botanical code, which is based on morphology and ecology, due to their perceived relationship to algae. However, this was challenged in the 1970's by Stanier and colleagues who advocated that since Cyanobacteria are bacteria, their nomenclature should be governed by the Bacteriological Code [45]. In 1979, Rippka and colleagues proposed to categorise Cyanobacteria into five subgroups based on cell organisation to simplify assignment for cultures and attempted to use generic nomenclature and definitions used by phycologists [46]. Section I and II are unicellular with section I reproducing by binary fission and section II reproducing by multiple fission. Section III, IV and V are filamentous. Section III and IV divide in one plane, whereas section V has the ability to form branching filaments. Section IV and V may also form akinetes or homogonia, with section V having a further ability to form branching filaments [47].

More recently, molecular methods have been used to classify Cyanobacteria, most notably the 16S rRNA genes and the internal transcribed spacer sequences (ITS) [47, 48]. Phylogenetic analysis of various protein coding sequences have also been used including the phycocyanin operon and its ITS [49], the β-subunit of RNA polymerase (RpoB) [50], ribulose bisphosphate carboxylase/oxygenase (RbcLX) [51] and other marker genes [38]. Another approach that has been used for cyanobacterial classification is the identification of synapomorphies that are specific for different phylogenetically

defined clades. In one study, >40 conserved indels were identified in proteins that were present in either all Cyanobacteria or its major clades [51]. Martin *et al.* (2003) identified 181 proteins in 7 out of 8 Cyanobacteria that were not found in other bacteria, supporting the monophyly of the phylum. However, at lower taxonomic ranks there is extensive discord between molecular and morphology-based classifications [52, 53]. Comparative analysis of whole genomes is likely to provide a definitive basis for classification of Cyanobacteria [53-60].

Recently, Shih and colleagues sequenced the genomes of 54 phylogenetically and phenotypically diverse cyanobacterial species enormously improving the genomic coverage of the phylum. They generated a species tree using a concatenation of 31 conserved proteins and grouped the Cyanobacteria into seven major subclades which they labelled A to G. This tree largely supports single gene-based classifications and contradicts morphology-based classification in numerous instances. These contradictions suggest that several morphological attributes such as filament cell type and the ability to form baeocyes evolved independently several times and are therefore not sound indicators of common ancestry in the Cyanobacteria [2].

1.3.1 Uncultured basal Cyanobacteria

In additional to cyanobacterial isolates, the 16S rRNA gene has been extensively applied to survey microbial communities without the need for microbial cultivation (*see Section 1.5*). Within the 16S rRNA-defined Cyanobacteria, two class-level groups of uncultured basal Cyanobacteria called 4C0d-2 and ML635J-21 have been identified in culture-independent studies [61, 62]. Four orders have been described for class 4C0d-2: YS2, MLE1-12, SM1D11 and SM2F09 [63-65]. Although members of this class have not been cultured, they have been identified in multiple 16S rRNA gene clone libraries and amplicon analyses. YS2 are the most widely reported and have been identified in mammal guts, including mice, yaks, gayals and swamp buffaloes [64, 66-70], human guts [71-73] and anaerobic digestors [74]. The order MLE1-12 has been found in rapeseed [63], deep marine sediments [75], drinking water [76], bioreactors [77] and arctic snow and meltwater [78]. SM1D11 have been detected in soil [65, 79, 80] and earthworm intestines [81] and SM2F09 representatives have been identified in hot springs and short-tailed shearwater but these data have not been published as yet.

Uncultivated microorganisms can also be characterised directly from environmental samples using shotgun sequencing (metagenomics; *see Section 1.5*). Recently, Di Rienzi and colleagues obtained six draft genomes [82] of members of class 4C0d-2 from metagenomes of human gut and aquifer samples. Importantly, these genomes did not contain any detectable photosynthetic genes. They reclassified class 4C0d-2 as a new phylum, Melainabacteria, based on their basal position relative to

photosynthetic Cyanobacteria and their lack of photosynthetic genes, categorising the representatives within two groups, environmental (non-gut) and gut [81]. However, the phylogenetic positioning of this lineage remains debatable as to whether it should be classified as a new phylum or a class within the Cyanobacteria.

1.4 Challenging the dogma that all Cyanobacteria are photosynthetic

The distribution of Melainabacteria includes many aphotic habitats and the study by Di Rienzi confirmed that genomes from this lineage lack photosynthesis genes (*see above*). If the Melainabacteria are confirmed to be part of the Cyanobacteria than the dogma that all Cyanobacteria are photosynthetic would be challenged. The fact that some Cyanobacteria are non-photosynthetic should not be controversial because all other phyla that contain photosynthetic representatives also contain non-photosynthetic representatives, except for the Chlorobi **(Figure 1.2)**. However, the Chlorobi form a robust monophyletic grouping with the non-photosynthetic phyla Bacteroidetes, Ignavibacteria and Marinimicrobia, so it could be contested whether the Chlorobi should include non-photosynthetic members if these were to be included in the phyla (**Figure 1.2**). Within the Proteobacteria and Firmicutes, photosynthetic representatives are sparsely spread suggesting either late acquisition(s) of photosynthesis or many independent losses [34]. Two of the seven classes of the phylum Chloroflexi, Anaerolineae and Chloroflexi, contain photosynthetic representatives and although there are only a few cultured Gemmatimonadetes, both photosynthetic [18] and non-photosynthetic [83] members have been reported. The phylum Acidobacteria, only has one recognised photosynthetic representative, *Candidatus* Chloroacidobacterium thermophilum [84].

1.5 Culture-independent molecular approaches

It is commonly cited that less than 1% of microorganisms in Nature are able to be obtained in pure culture leading to the conclusion that microbiology has been limited by a "great plate-count anomaly" [85]. Over the past four decades a plethora of molecular tools have been developed allowing us to bypass this cultivation bottleneck, resulting in the discovery of new organisms.

1.5.1 Community profiling using 16S rRNA genes

The use of the universally conserved 16S rRNA gene to classify microorganisms was pioneered by Carl Woese (1977) providing the first objective evolutionary framework for microbial taxonomy. In the mid-1980's Norm Pace and colleagues developed methods to use the 16S rRNA gene to identify microorganisms in environmental samples without the need for cultivation, bypassing the cultivation bottleneck. These pioneering advances have revolutionised our understanding of

microbial evolution and ecology [42, 87, 88]. PCR-based fingerprinting methods that target the 16S rRNA gene, including Terminal Restriction Fragment Length Polymorphism (T-RFLP) [89], Amplified Ribosomal DNA Restriction Analysis (ARDRA) [90], Denaturing Gradient Gel Electrophoresis (DGGE) [91] and Automated Ribosomal Intergenic Spacer Analysis (ARISA) [92] are laborious, costly and time-consuming, limiting the number of samples that can be processed.

Recently a number of high-throughput DNA sequencing technologies (or next-generation sequencing) has become available, such as the 454 GS FLX and GS Junior (Roche), MiSeq and HiSeq2000 (Illumina), SOLiD (Applied Biosystems/Life-Technologies), IonTorrent PGM (Life Technologies) and the PacBio RS from Pacific Biosciences [93-96]. These platforms provide large amounts of sequence data and many have been applied to sequencing the 16S rRNA gene to profile microbial communities. They are less time consuming, more cost effective and have much greater resolution than traditional fingerprinting methods. A large number of samples can be run in parallel by multiplexing and samples can be split based on unique sample-specific barcodes [97]. Although the use of the 16S rRNA gene has led to the discovery many new microbial lineages, the data is only able to provide a profile of the community (membership) and does not give insight into the genetics, physiology and biochemistry of the members [98].

1.5.2 Metagenomics

Metagenomics, here defined as the shotgun (random) sequencing and analysis of bulk DNA extracted from environmental samples can be used to determine both the identity and putative function of a microbial community [99]. Basic bioinformatic steps in metagenomics are read trimming and assembly, which produce contigs (overlapping sequence data (reads)), binning (grouping contigs together based on high statistical support), annotation and comparative analysis, which are described in turn below (**Figure 1.3**). Gene absence and partial pathways can be difficult to interpret but overall metagenomics can be useful for developing physiological hypotheses and testing evolutionary ideas. Metagenomics was first used to identify gene families across the entire microbial community, a process termed gene-centric analysis (GCA), due to initial difficulties in extracting genomes from individual populations. GCA identifies different relative abundances of gene families between microbial communities, thus providing clues as to important functionalities for a given habitat [100]. With increasing sequencing depth and computational power, it is now possible to separate (bin) near complete genomes from bulk metagenomic data to explore the metabolic potential of individual populations.

1.5.2.1 Read trimming and assembly

Short-insert paired-end read sequences (reads with <1kb inserts) are typically trimmed using tools such as cutadapt [101] or Trimmomatic [102] so that only high quality sequences are used for metagenome assembly. The trimmed reads are assembled into contiguous sequences called contigs using a metagenome assembler such as MetaVelvet [103], Ray Meta [104], Minimus [105], Newbler or CLC Genomics Workbench (http://www.clcbio.com). MetaVelvet and Ray Meta use a de Bruijn graph algorithm, in which the sequencing reads are cut into shorter k-mers (DNA sequences consisting of a fixed number (k) of bases) which form the de Bruijn graph and infer the genome sequence. Both Minimus and Newbler use an Overlap-Layout-Consensus method, where the assembler overlaps all the reads, carries out a layout of all reads and infers a consensus sequence [106]. CLC Genomics Workbench is a popular commercial assembler with its own proprietary algorithm but is likely a de Bruijn graph assembler. Contigs can be joined using paired end or mate pair information (links), a process termed scaffolding. Scaffolding tools such as SSPACE (SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension) [107] or Bambus 2 [108] scaffold contigs by mapping long-insert mate-pair reads (reads with 2-5kb inserts) to the contigs, potentially resolving repetitive structures. Both SSPACE and Bambus use a greedy algorithm either joining together contigs with the most links first and ignoring subsequent edges that conflict with an existing join or building the first scaffold from the longest contig and continuing to make joins as long as the majority of read pairs support the join [109].

1.5.2.2 Binning

Contigs and scaffolds can be grouped (binned) together into component populations based on a number of features including similarity/homology to reference genomes [110, 111], sequence composition, e.g. GC content [112, 113], and sequencing coverage in single or multiple samples [114, 115]. MEGAN (Metagenome Analyzer) [110] is a homology-based binning method that compares scaffolds against a database of known sequences. It then estimates and explores the content of the dataset to summarise and order the results and uses a simple algorithm to assign each read to the lowest common ancestor. CARMA [111] is a phylogenetic algorithm that searches for conserved Pfam (protein family) domains and protein families in unassembled sequencing reads. The gene fragments are classified based on reconstruction of phylogenetic trees of each matching Pfam family. Studies on cultivated organisms have shown that each species has its own unique nucleotide composition. The frequency at which these nucleotides occur are unique between species but are conserved throughout the genome [116]. Genome signatures in the form of k-mer frequencies (most commonly tetramers) can be used to bin scaffolds into their component genomes. Two examples of methods that use sequence composition-based binning is PhyloPythia [112] and

Emergent self-organising maps (ESOMs) [113]. Phylophythia uses the relative frequencies of 4 to 6-mers as features to train support vector machine classifiers, which then assign the query sequence to a bin. ESOMs are unsupervised neural network algorithms that cluster sequencing fragments (contigs that are cut into fragments) containing tetranucleotides into two-dimensional borderless maps. The sequencing fragments that cluster closer together are more similar and are potentially from the same population genome [113]. Read depth of assembled contigs (coverage) has been used extensively to group contigs together which is based on the fact that higher relative abundance populations are represented by more reads in a metagenome than lower abundance populations [117]. A recent modification of this approach that is gaining popularity uses coverage information from multiple related metagenomes to create a coverage profile for a given population. This approach termed differential coverage binning, is capable of recovering low abundance populations (<1% relative abundance) given enough sequencing depth [114]. An automated binning tool based primarily on differential coverage binning, GroopM [115], has recently been described and requires a minimum of three related metagenomes to bin populations.

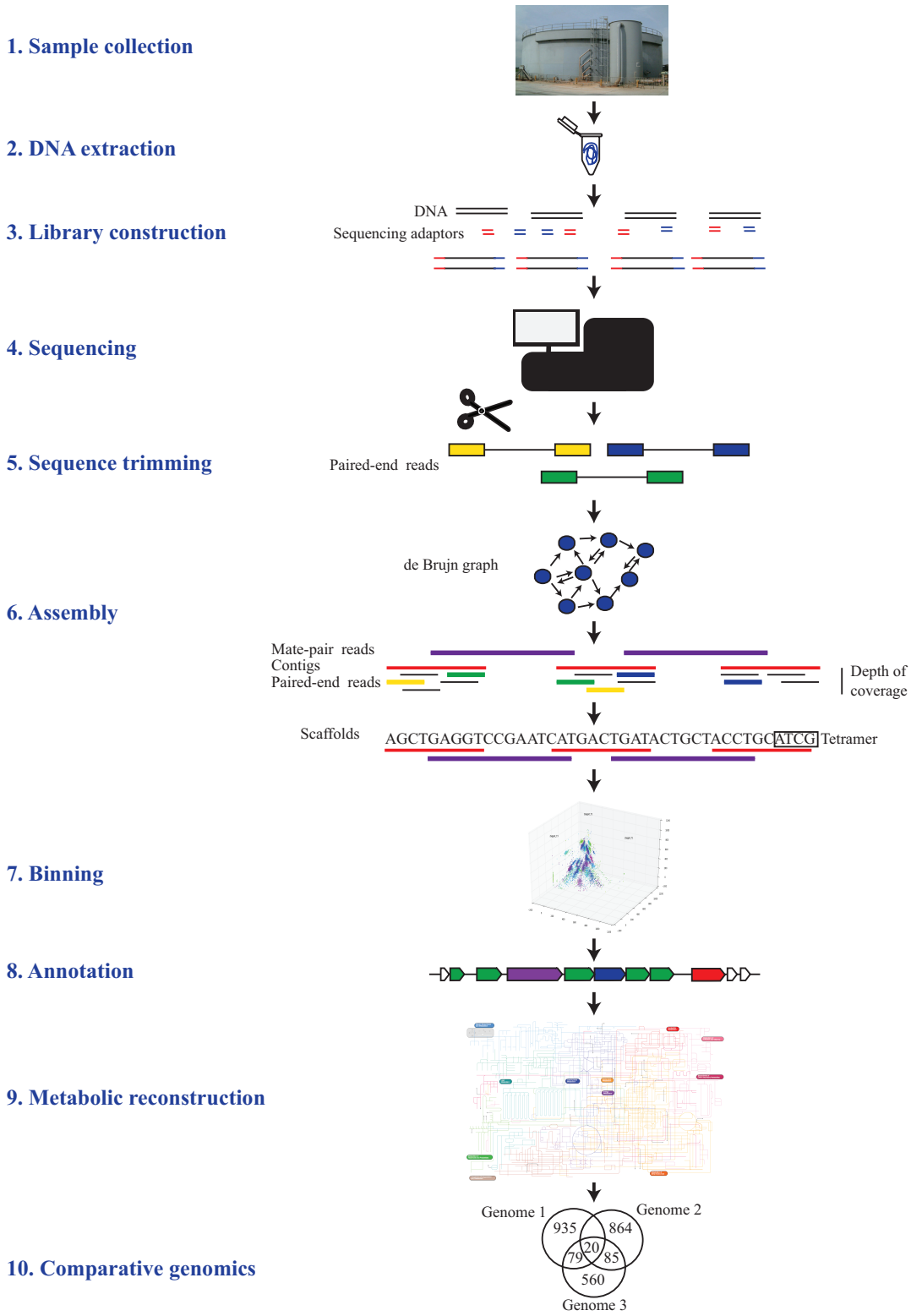
1.5.2.3 Annotation (gene calling)

Putative genes, ribosomal and transfer RNAs (rRNAs and tRNAs respectively) encoded on the binned scaffolds can be obtained by annotation, the process of calling open reading frames (ORFs) and assigning function using a database of characterised genes. Bacterial genome annotation can be processed by annotation tools such as DIYA (Do-It-Yourself Annotator) [118], the Integrated Microbial Genomes (IMG) system [119], RAST (Rapid Annotation using Subsystem Technology) [120] or Prokka [121]. These annotation tools rely on gene recognition software, such as Glimmer (Gene Locator and Interpolated Markov ModelER) [122], GeneMark [123] or Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) [124] that identify coding regions using different algorithms. The annotated data can be mapped onto KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, which can be used for biological interpretation of higher-level functions. In addition, the MetaCyc Database [125] can be used to identify metabolic pathways used by an organism that may not be present in the KEGG database. Specialist databases can also be used to identify specific attributes within a genome, for example the CAZy (Carbohydrate-Active enZYmes) database describes the families of structurally-related catalytic and carbohydrate-binding molecules of enzymes, and TransportDB [126] is a database of cytoplasmic membrane transport systems and outer membrane channels. The genome annotations, KEGG maps and specialist databases, can be used to reconstruct the overall metabolism available to a given organism (genome).

1.5.2.4 Comparative analysis

Comparative genomics, in which the genomic features of different organisms are compared, can be used to identify sequences that share a common ancestry called orthologous sequences (orthologs), paralogs (duplicated genes) and co-orthologs (sequences that share a common ancestry and have been duplicated). Orthologs are of interest because it is expected that microbes maintain at least part of their (ancestral) biological function [127]. Tools such as OrthoMCL-DB [128], Ensembl [129], Proteinortho [127] and eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) [130] can be used to identify orthologs, paralogs and co-orthologs through reciprocal best alignment heuristics to protein sequences found in multiple genomes of interest. The results from these analyses can be used to describe the pan-genome of a given set of genomes, which comprises the core genome (genes that are shared by all genomes being compared), accessory genome (genes that are present in two or more genomes) and unique genes (genes that are only present in one genome) [131]. The core genome typically includes genes responsible for core functions (e.g. transcription, translation), whereas the accessory genome usually encodes genes that confer a selective advantage to a given species or mobilisable elements [132]. Such comparative analyses have been used extensively on microbial isolate genomes [132-134], but are increasingly being applied on sets of population genomes derived from metagenomic datasets [135].

Figure 1.3. Overview of binning a bacterial genome from a metagenomic dataset and constructing a metabolic schema



1.5.3 Metatranscriptomics

Metagenomics provides only the potential functionality of a community or population, as it interrogates the genomic blueprints and not the expressed products of those blueprints [136]. One bioinformatic proxy, predicted highly expressed (PHX) gene analysis can be used to predict gene expression levels based on codon usage differences [137]. However, gene expression data is preferable as it directly confirms that products are formed and provides context-dependent information on gene expression, such as expression at different pH or temperatures. "Omics" approaches that provide information on gene expression include metaproteomics which identifies proteins present in microbial communities by referencing peptides to metagenomic datasets [138] and metabolomics which identifies some of the collection of small molecules produced by cells [139]. However, the most widely used expression based omic technology is metatranscriptomics which identifies which genes encoded in a metagenome are transcribed and therefore which metabolic pathways are active [140]. Metatranscriptomics provides a snapshot of the total RNA present in a microbial community at a given time and comprises coding (messenger) RNA (mRNA) and non-coding RNA (rRNA, tRNA, regulatory RNA and other RNA species) [141]. A typical metatranscriptomic workflow is shown in **Figure 1.4**, with the major steps described below in more detail.

1.5.3.1 Sample collection and RNA extraction

Environmental samples collected for RNA extraction should be snap-frozen immediately and stored in an RNA preserving buffer (e.g. Lifeguard), as mRNAs have a half-life typically ranging from a fraction of a minute to hours [142-144]. During RNA extraction, DNA is degraded with DNAse and an inhibitor of RNases can be added to avoid RNA degradation. RNA can be extracted using several methods including guanidinium thiocyanate-phenol-chloroform extraction [145] or standard laboratory kits such as NucleoSpin RNA (Macherey-Nagel) or the RNA PowerSoil Total RNA Isolation Kit (Mo Bio) which use either guanidinium thiocyanate or a bead-beating method with a phenol-chloroform extraction respectively.

1.5.3.2 Enrichment of mRNA, cDNA synthesis and high throughput sequencing

Messenger RNA only contributes 1-5% of the total RNA in a typical bacterial cell, with rRNA comprising most of the remainder (up to 90% of total RNA) [146]. Therefore, rRNAs are typically depleted prior to sequencing, for example, through subtractive hybridisation using antisense rRNA probes bound to magnetic beads that hybridise to conserved regions of rRNA molecules and remove them by drawing down the beads [147] or by using exonucleases that preferentially digest rRNA by targeting 5'-monophosphate ends [148]. Once the rRNA is depleted, the remaining RNA

is fragmented and used to synthesise single-stranded complementary DNA (cDNA) by reverse transcription with random hexamers. A second round of PCR is used to amplify the second strand of cDNA and to attach sample-specific barcodes which allows multiple samples to be pooled and sequenced together and bioinformatically separated during data analysis. The pooled cDNAs are purified prior to sequencing (*see Section 1.5.1*). Messenger RNA is strand-specific and unidirectional, therefore terminal tagging at each PCR step can be used to identify which strand corresponds to the mRNA strand. This also allows for detection of contaminating DNA, as these sequences will be bi-directional.

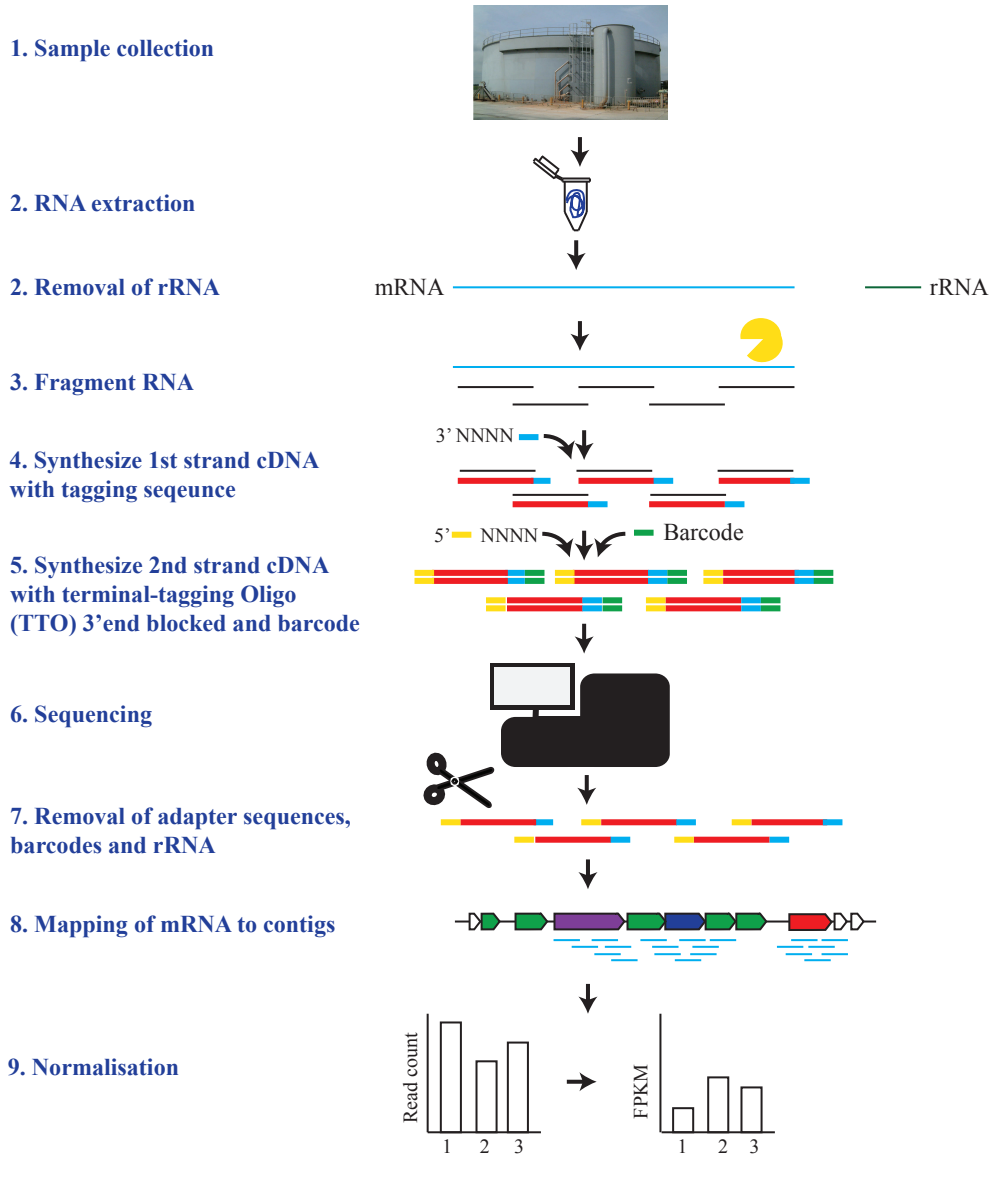
1.5.3.3 Metatranscriptomic data analysis

As for the metagenomic sequencing reads, the metatranscriptomic reads are trimmed so that only high quality sequences are used and sequencing adaptors are removed. Paired-end reads are merged into single longer reads and any residual rRNA can be removed using tools such as rRNASelector, which uses a Hidden Markov Model (HMM) to sort reads against a pre-built database [149]. Another tool for removing rRNA reads is SortMeRNA [146], which uses an algorithm to filter sequencing reads that are similar to a user-provided set of rRNA sequences. There are two main ways typically used to reconstruct a transcriptome: 'genome-guided' and 'genome-independent'. Genome-guided methods rely on mapping cDNA reads to a reference genome or metagenome, whereas genome-independent methods *de novo* assemble reads into transcripts without the use of a reference [150]. In the genome-guided method, the trimmed RNA paired-end reads are mapped back to the genomic contigs to identify which genes are being expressed. Genome-guided alignment can be achieved using a seed method, such as MAQ (mapping and assembly with quality) or Stampy, or the Burrows-Wheeler transform method with alignment tools such as BWA (Burrows-Wheeler Aligner; [151]), Bowtie2 [152] or SOAP3-dp (Short Oligonucleotide Analysis Package; [153]). Seed methods find matches for short subsequences, called 'seeds', where it is assumed that at least one seed will perfectly match the reference. Each seed is used to find an area that closely matches and then more sensitive methods can be used to extend seeds to full alignments. In the Burrows-Wheeler transform methods, the genomes are compacted into a data structure which is efficient for searching for perfect matches but becomes slower as it allows for mismatches [150]. Genome-guided and -independent assemblies can be visualised using tools such as Geneious [154] or Tablet [155] to confirm that reads are unidirectional (usually bi-directional paired reads indicate DNA contamination) and the level of transcription based on read coverage.

1.5.3.4 Differential gene expression analysis

Read coverage is standardly normalised to allow unbiased comparisons of metatranscriptomic data. Normalisation enables accurate comparisons between and within samples and adjusts for systematic and technical biases, such as gene length and GC-content [156]. Fragments (paired-end sequences) per kilobase of transcript per million mapped (FPKM) is one of the most used methods for normalisation. It takes into account the length and total number of mapped fragments in a sample as more reads will map to larger fragments than shorter fragments of the same abundance, skewing expression levels. Other methods include total count, upper quartile and median of gene counts and Trimmed Mean of M-values (TMM), in which a TMM factor is computed for each lane with one lane used as a reference [156].

Figure 1.4. Overview of genome-guided transcriptomics



Modified from ScriptSeq manual

1.6 Summary of chapters

Chapter Two describes the sequencing and analysis of five near-complete Melainabacteria population genomes. Environmental samples containing Melainabacteria representatives were identified by 16S rRNA community profiling and metagenomes were prepared from these samples. Melainabacteria population genomes were recovered from these samples using differential coverage binning which were then used for phylogeny and comparative analyses. From these analyses it was proposed that the Melainabacteria is a class within the phylogenetically defined Cyanobacteria and not a sister phylum as previously concluded. Four new orders within the class Melainabacteria were also proposed based on the population genomes. During analysis of Melainabacteria 16S rRNA genes, the sequence of a putative cultured representative, *Vampirovibrio chlorellavorus,* was discovered. Chapter Three describes the sequencing and comparative analysis of a near-complete draft genome of this predatory bacterium obtained directly from 36 year-old lyophilised cells co-cultured with its host. A detailed schema of how *V. chlorellavorus* attaches and attacks its microalgae prey, *Chlorella vulgaris*, is proposed. In Chapter Four, two Melainabacteria population genomes belonging to the order Obscuribacterales were recovered from a Swedish permafrost sample primarily via sequence composition and coverage analysis. Metabolic reconstruction of the two genomes was used to identify potential adaptations for living in the cold environment. Metatranscriptomics data from the same samples was used to identify which genes are expressed in the two genomes. Chapter Five summarises the findings in this thesis and suggests future directions for understanding the Melainabacteria.

1.7 References

1. Martin, W., et al., Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(19): p. 12246-51.
2. Shih, P.M., et al., Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(3): p. 1053-8.
3. Raines, C., The Calvin cycle revisited. Photosynthesis Research, 2003. **75**(1): p. 1-10.
4. Hess, W.R., Cyanobacterial genomics for ecology and biotechnology. Current Opinion in Microbiology, 2011. **14**(5): p. 608-14.
5. Cohen, Y., et al., Sulphide-dependent anoxygenic photosynthesis in the cyanobacterium *Oscillatoria limnetica*. Nature, 1975. **257**(5526): p. 489-492.

6. Thompson, A.W., et al., Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. Science, 2012. **337**(6101): p. 1546-1550.
7. Zehr, J.P., et al., Globally Distributed Uncultivated Oceanic N_2-Fixing Cyanobacteria Lack Oxygenic Photosystem II. Science, 2008. **322**(5904): p. 1110-1112.
8. Zurawell, R.W., et al., Hepatotoxic Cyanobacteria: A Review of the Biological Importance of Microcystins in Freshwater Environments. Journal of Toxicology and Environmental Health, Part B, 2005. **8**(1): p. 1-37.
9. Rocap, G., et al., Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature, 2003. **424**(6952): p. 1042-1047.
10. Garcia-Pichel, F., A. López-Cortés, and U. Nübel, Phylogenetic and Morphological Diversity of Cyanobacteria in Soil Desert Crusts from the Colorado Plateau. Applied and Environmental Microbiology, 2001. **67**(4): p. 1902-1910.
11. Warren-Rhodes, K., et al., Hypolithic Cyanobacteria, Dry Limit of Photosynthesis, and Microbial Ecology in the Hyperarid Atacama Desert. Microbial Ecology, 2006. **52**(3): p. 389-398.
12. Kauff, F. and B. Büdel, Phylogeny of cyanobacteria: an overview, in Progress in Botany 72. 2011, Springer. p. 209-224.
13. Hoiczyk, E. and A. Hansel, Cyanobacterial Cell Walls: News from an Unusual Prokaryotic Envelope. Journal of Bacteriology, 2000. **182**(5): p. 1191-1199.
14. Dagan, T., et al., Genomes of Stigonematalean Cyanobacteria (Subsection V) and the Evolution of Oxygenic Photosynthesis from Prokaryotes to Plastids. Genome Biology and Evolution, 2013. **5**(1): p. 31-44.
15. Griese, M., C. Lange, and J. Soppa, Ploidy in cyanobacteria. FEMS Microbiology Letters, 2011. **323**(2): p. 124-131.
16. Blankenship, R., Origin and early evolution of photosynthesis. Photosynthesis Research, 1992. **33**(2): p. 91-111.
17. Hohmann-Marriott, M.F. and R.E. Blankenship, Evolution of Photosynthesis. Annual Review of Plant Biology, 2011. **62**(1): p. 515-548.
18. Zeng, Y., et al., Functional type 2 photosynthetic reaction centers found in the rare bacterial phylum Gemmatimonadetes. Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(21): p. 7795-7800.
19. Xiong, J., Photosynthesis: what color was its origin? Genome Biology, 2006. **7**(12): p. 245.
20. Xiong, J. and C.E. Bauer, Complex evolution of photosynthesis. Annual Review of Plant Biology, 2002. **53**: p. 503-21.

21. Olson, J.M. and R.E. Blankenship, Thinking about the evolution of photosynthesis. Photosynthesis Research, 2004. **80**(1-3): p. 373-86.
22. Schidlowski, M., A 3,800-million-year isotopic record of life from carbon in sedimentary rocks. Nature, 1988. **333**(6171): p. 313-18.
23. McCollom, T.M. and J.S. Seewald, Carbon isotope composition of organic compounds produced by abiotic synthesis under hydrothermal conditions, 2006. **243**(1-2): p. 74-84.
24. Des Marais, D.J., When Did Photosynthesis Emerge on Earth? Science, 2000. **289**(5485): p. 1703-1705.
25. Hofmann, H.J. and G.D. Jackson, Proterozoic minstromatolites with radial-fibrous fabric. Sedimentology, 1987. 34(6): p.963-71.
26. Schopf, J.W. Microfossils of the Early Archean Apex Chert: New Evidence of the Antiquity of Life, 1993. **260**(5108): p. 640-46.
27. Javaux, E.J., Marshall, C.P. and A. Bekker, Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. Nature, 2010. **463**:p. 934-38.
28. Summons, R. E., Jahnke, L. L., Hope, J. M. and G.A. Logan, 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic phostosynthesis. Nature, 1999. **400**: p.554-57.
29. Brocks, J.J., Logan, G.A., Buick, R. and R.E. Summons, Archean Molecular Fossils and the Early Rise of Eukaryotes. Science, 1999. **285**(5430): p.1033-36.
30. French, K.L., Hallmann, C., Hope, J.M., Schoon, P.L., Zumberge, J.A., Hoshino, Y., Peters, C.A., George, S.C., Love, G.D., Brocks, J.J., Buick, R. and R.E. Summons, Reappraisal of hydrocarbon biomarkers in Archean rocks. Proceedings of the Naitonal Academy of Sciences of the United States of America, 2015. **112**(19): p5915-20.
31. Tice, M.M. and D.R. Lowe, Photosynthetic microbial mats in the 3,416-Myr-old ocean. Nature, 2004. **431**(7008): p. 549-552.
32. Johnson, J.E., Gerpheide, A., Lamb, M.P. and W.W. Fischer, O2 constraints from Paleoproterozoic detrital pyrite and uraninite. Geological Society of America Bulletin, 2014. **126**(5-6): p. 813-30.
33. Raymond, J. and R.E. Blankenship, Biosynthetic pathways, gene replacement and the antiquity of life. Geobiology, 2004. **2**(4): p. 199-203.
34. Sousa, F.L., et al., Chlorophyll biosynthesis gene evolution indicates photosystem gene duplication, not photosystem merger, at the origin of oxygenic photosynthesis. Genome Biology and Evolution, 2013. **5**(1): p. 200-216.
35. Larkum, A.D., The Evolution of Chlorophylls and Photosynthesis, in Chlorophylls and Bacteriochlorophylls, B. Grimm, et al., Editors. 2006, Springer Netherlands. p. 261-282.

36. Cardona, T., A fresh look at the evolution and diversification of photochemical reaction centers. Photosynthesis Research, 2014: p. 1-24.
37. Xiong, J., et al., Molecular Evidence for the Early Evolution of Photosynthesis. Science, 2000. **289**(5485): p. 1724-1730.
38. Gupta, R.S., Evolutionary relationships among photosynthetic bacteria. Photosynthesis Research, 2003. **76**(1-3): p. 173-83.
39. Raymond, J., et al., Whole-genome analysis of photosynthetic prokaryotes. Science, 2002. **298**(5598): p. 1616-20.
40. Granick, S., Evolution of Heme and Chlorophyll, in Evolving Genes and Proteins, V. Bryson and H.J. Vogel, Editors. 1965, Academic Press. p. 67-88.
41. Shi, T., et al., Protein Interactions Limit the Rate of Evolution of Photosynthetic Genes in Cyanobacteria. Molecular Biology and Evolution, 2005. **22**(11): p. 2179-2189.
42. Hugenholtz, P., B.M. Goebel, and N.R. Pace, Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology, 1998. **180**(18): p. 4765-74.
43. Mulkidjanian, A.Y., et al., The cyanobacterial genome core and the origin of photosynthesis. Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(35): p. 13126-13131.
44. Stanier, R.Y. and G. Cohen-Bazire, Phototrophic prokaryotes: the cyanobacteria. Annual Review of Microbiology, 1977. **31**: p. 225-74.
45. Wilmotte, A., Molecular Evolution and Taxonomy of the Cyanobacteria, in The Molecular Biology of Cyanobacteria, D. Bryant, Editor. 2004, Springer Netherlands. p. 1-25.
46. Rippka, R., et al., Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. Journal of General Microbiology, 1979. **111**(1): p. 1-61.
47. Honda, D., A. Yokota, and J. Sugiyama, Detection of Seven Major Evolutionary Lineages in Cyanobacteria Based on the 16S rRNA Gene Sequence Analysis with New Sequences of Five Marine Synechococcus Strains. Journal of Molecular Evolution, 1999. **48**(6): p. 723-739.
48. Rocap, G., et al., Resolution of *Prochlorococcus* and *Synechococcus* Ecotypes by Using 16S-23S Ribosomal DNA Internal Transcribed Spacer Sequences. Applied and Environmental Microbiology, 2002. **68**(3): p. 1180-1191.
49. Berrendero, E., E. Perona, and P. Mateo, Genetic and morphological characterization of *Rivularia* and *Calothrix* (Nostocales, Cyanobacteria) from running water. International Journal of Systematic and Evolutionary Microbiology, 2008. **58**(2): p. 447-460.

50. Rajaniemi, P., et al., Phylogenetic and morphological evaluation of the genera *Anabaena*, *Aphanizomenon*, *Trichormus* and *Nostoc* (Nostocales, Cyanobacteria). International Journal of Systematic and Evolutionary Microbiology, 2005. **55**(1): p. 11-26.
51. Gupta, R.S., Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. International Journal of Systematic and Evolutionary Microbiology, 2009. **59**(10): p. 2510-2526.
52. Martin, K.A., et al., Cyanobacterial signature genes. Photosynth Res, 2003. **75**(3): p. 211-21.
53. Schirrmeister, B., A. Antonelli, and H. Bagheri, The origin of multicellularity in cyanobacteria. BMC Evolutionary Biology, 2011. **11**(1): p. 45.
54. SÁNchez-Baracaldo, P., P.K. Hayes, and C.E. Blank, Morphological and habitat evolution in the Cyanobacteria using a compartmentalization approach. Geobiology, 2005. **3**(3): p. 145-165.
55. Ciccarelli, F.D., et al., Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science, 2006. **311**(5765): p. 1283-1287.
56. Zhaxybayeva, O., et al., Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. Genome Research, 2006. **16**(9): p. 1099-1108.
57. Shi, T. and P.G. Falkowski, Genome evolution in cyanobacteria: the stable core and the variable shell. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(7): p. 2510-5.
58. Swingley, W.D., R.E. Blankenship, and J. Raymond, Integrating Markov Clustering and Molecular Phylogenetics to Reconstruct the Cyanobacterial Species Tree from Conserved Protein Families. Molecular Biology and Evolution, 2008. **25**(4): p. 643-654.
59. Larsson, J., J. Nylander, and B. Bergman, Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evolutionary Biology, 2011. **11**(1): p. 187.
60. Beck, C., et al., The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. BMC Genomics, 2012. **13**(1): p. 56.
61. McDonald, D., et al., An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J, 2012. **6**(3): p. 610-8.
62. Quast, C., et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 2013. **41**(D1): p. D590-D596.
63. Kaiser, O., A. Puhler, and W. Selbitschka, Phylogenetic Analysis of Microbial Diversity in the Rhizoplane of Oilseed Rape (*Brassica napus* cv. Westar) Employing Cultivation-Dependent and Cultivation-Independent Approaches. Microbial Ecology, 2001. **42**(2): p. 136-149.

64. Ley, R.E., et al., Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(31): p. 11070-11075.
65. Elshahed, M.S., et al., Novelty and Uniqueness Patterns of Rare Members of the Soil Biosphere. Applied and Environmental Microbiology, 2008. **74**(17): p. 5422-5428.
66. Ley, R.E., et al., Microbial ecology: Human gut microbes associated with obesity. Nature, 2006. **444**(7122): p. 1022-1023.
67. Stecher, B., et al., *Salmonella enterica* serovar typhimurium exploits inflammation to compete with the intestinal microbiota. PLoS Biol, 2007. **5**(10): p. 2177-89.
68. Tajima, K., et al., Influence of high temperature and humidity on rumen bacterial diversity in Holstein heifers. Anaerobe, 2007. **13**(2): p. 57-64.
69. Monteils, V., et al., Potential core species and satellite species in the bacterial community within the rabbit caecum. FEMS Microbiology Ecology, 2008. **66**(3): p. 620-9.
70. Yang, S., et al., Bacterial diversity in the rumen of Gayals (*Bos frontalis*), Swamp buffaloes (*Bubalus bubalis*) and Holstein cow as revealed by cloned 16S rRNA gene sequences. Molecular Biology Reports, 2010. **37**(4): p. 2063-73.
71. Dethlefsen, L., et al., The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing. PLoS Biology, 2008. **6**(11): p. e280.
72. Ley, R.E., et al., Evolution of Mammals and Their Gut Microbes. Science, 2008. **320**(5883): p. 1647-1651.
73. Turnbaugh, P.J., et al., A core gut microbiome in obese and lean twins. Nature, 2009. **457**(7228): p. 480-4.
74. Riviere, D., et al., Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. ISME J, 2009. **3**(6): p. 700-714.
75. Reed, D.W., et al., Microbial Communities from Methane Hydrate-Bearing Deep Marine Sediments in a Forearc Basin. Applied and Environmental Microbiology, 2002. **68**(8): p. 3759-3770.
76. Williams, M.M., et al., Phylogenetic diversity of drinking water bacteria in a distribution system simulator. Journal of Applied Microbiology, 2004. **96**(5): p. 954-64.
77. Liu, B., et al., *Thauera* and *Azoarcus* as functionally important genera in a denitrifying quinoline-removal bioreactor as revealed by microbial community structure comparison. FEMS Microbiology Ecology, 2006. **55**(2): p. 274-86.
78. Larose, C., et al., Microbial sequences retrieved from environmental samples from seasonal Arctic snow and meltwater from Svalbard, Norway. Extremophiles, 2010. **14**(2): p. 205-212.
79. Lesaulnier, C., et al., Elevated atmospheric CO_2 affects soil microbial diversity associated with trembling aspen. Environmental Microbiology, 2008. **10**(4): p. 926-41.

80. Cruz-Martinez, K., et al., Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. ISME J, 2009. **3**(6): p. 738-44.
81. Rattray, R.M., et al., Microbiomic comparison of the intestine of the earthworm *Eisenia fetida* fed ergovaline. Current Microbiology, 2010. **60**(3): p. 229-35.
82. Di Rienzi, S.C., et al., The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. eLife, 2013. **2**: p. e01102.
83. Zhang, H., et al., *Gemmatimonas aurantiaca* gen. nov., sp. nov., a Gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov. International Journal of Systematic and Evolutionary Microbiology, 2003. **53**(4): p. 1155-1163.
84. Bryant, D.A., et al., *Candidatus* Chloracidobacterium thermophilum: An Aerobic Phototrophic Acidobacterium. Science, 2007. **317**(5837): p. 523-526.
85. Staley, J.T. and A. Konopka, Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annual Review of Microbiology, 1985. **39**: p. 321-46.
86. Woese, C.R. and G.E. Fox, Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America, 1977. **74**(11): p. 5088-5090.
87. Olsen, G.J., et al., Microbial Ecology and Evolution: A Ribosomal RNA Approach. Annual Review of Microbiology, 1986. **40**(1): p. 337-365.
88. Tringe, S.G. and P. Hugenholtz, A renaissance for the pioneering 16S rRNA gene. Current Opinion in Microbiology, 2008. **11**(5): p. 442-446.
89. Osborn, A.M., E.R. Moore, and K.N. Timmis, An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. Environmental Microbiology, 2000. **2**(1): p. 39-50.
90. De Baere, T., et al., Evaluation of amplified rDNA restriction analysis (ARDRA) for the identification of cultured mycobacteria in a diagnostic laboratory. BMC Microbiology, 2002. **2**(1): p. 4.
91. Muyzer, G., DGGE/TGGE a method for identifying genes from natural ecosystems. Current Opinion in Microbiology, 1999. **2**(3): p. 317-322.
92. Fisher, M.M. and E.W. Triplett, Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. Applied and Environmental Microbiology, 1999. **65**(10): p. 4630-6.

93. Caporaso, J.G., et al., Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J, 2012. **6**(8): p. 1621-1624.
94. Carneiro, M.O., et al., Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics, 2012. **13**: p. 375.
95. Jünemann, S., et al., Bacterial Community Shift in Treated Periodontitis Patients Revealed by Ion Torrent 16S rRNA Gene Amplicon Sequencing. PLoS ONE, 2012. **7**(8): p. e41606.
96. Pilloni, G., et al., Testing the Limits of 454 Pyrotag Sequencing: Reproducibility, Quantitative Assessment and Comparison to T-RFLP Fingerprinting of Aquifer Microbes. PLoS ONE, 2012. **7**(7): p. e40467.
97. Sogin, M.L., et al., Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(32): p. 12115-12120.
98. Handelsman, J., Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiology and Molecular Biology Reviews, 2004. **68**(4): p. 669-685.
99. Hugenholtz, P. and G.W. Tyson, Microbiology: Metagenomics. Nature, 2008. **455**(7212): p. 481-483.
100. Tringe, S.G., et al., Comparative Metagenomics of Microbial Communities. Science, 2005. **308**(5721): p. 554-557.
101. Martin, M., Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 2011. **17**(1).
102. Bolger, A.M., M. Lohse, and B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014. **30**(15): p. 2114-2120.
103. Namiki, T., et al., MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Research, 2012. **40**(20): p. e155.
104. Boisvert, S., et al., Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biology, 2012. **13**(12): p. R122.
105. Sommer, D., et al., Minimus: a fast, lightweight genome assembler. BMC Bioinformatics, 2007. **8**(1): p. 64.
106. Li, Z., et al., Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Briefings in Functional Genomics, 2012. **11**(1): p. 25-37.
107. Boetzer, M., et al., Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 2011. **27**(4): p. 578-9.
108. Koren, S., T.J. Treangen, and M. Pop, Bambus 2: scaffolding metagenomes. Bioinformatics, 2011. **27**(21): p. 2964-2971.

109. Hunt, M., et al., A comprehensive evaluation of assembly scaffolding tools. Genome Biology, 2014. **15**(3): p. R42.
110. Huson, D.H., et al., MEGAN analysis of metagenomic data. Genome Research, 2007. **17**(3): p. 377-386.
111. Krause, L., et al., Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Research, 2008. **36**: p. 2230 - 2239.
112. McHardy, A.C., et al., Accurate phylogenetic classification of variable-length DNA fragments. Nature Methods, 2007. **4**(1): p. 63-72.
113. Dick, G., et al., Community-wide analysis of microbial genome sequence signatures. Genome Biology, 2009. **10**(8): p. R85.
114. Albertsen, M., et al., Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature Biotechnology, 2013. **31**(6): p. 533-538.
115. Imelfort, M., et al., GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ, 2014. **2**: p. e603.
116. Karlin, S., J. Mrázek, and A.M. Campbell, Compositional biases of bacterial genomes and evolutionary implications. Journal of Bacteriology, 1997. **179**(12): p. 3899-913.
117. Tyson, G.W., et al., Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, 2004. **428**: p. 37 - 43.
118. Stewart, A.C., B. Osborne, and T.D. Read, DIYA: a bacterial annotation pipeline for any genomics lab. Bioinformatics, 2009. **25**(7): p. 962-963.
119. Markowitz, V.M., et al., IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Research, 2012. **40**(D1): p. D115-D122.
120. Overbeek, R., et al., The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research, 2014. **42**: p. D206-D214.
121. Seemann, T., Prokka: rapid prokaryotic genome annotation. Bioinformatics, 2014.
122. Delcher, A.L., et al., Improved microbial gene identification with GLIMMER. Nucleic Acids Research, 1999. **27**(23): p. 4636-4641.
123. Besemer, J. and M. Borodovsky, GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research, 2005. **33**: p. W451-W454.
124. Hyatt, D., et al., Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 2010. **11**(1): p. 119.
125. Caspi, R., et al., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research, 2012. **40**: p. D742-D753.

126. Ren, Q., K. Chen, and I.T. Paulsen, TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Research, 2007. **35**: p. D274-D279.
127. Lechner, M., et al., Proteinortho: Detection of (Co-)orthologs in large-scale analysis. BMC Bioinformatics, 2011. **12**(1): p. 124.
128. Fischer, S., et al., Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups, in Current Protocols in Bioinformatics. 2002, John Wiley & Sons, Inc.
129. Hubbard, T., et al., The Ensembl genome database project. Nucleic Acids Research, 2002. **30**(1): p. 38-41.
130. Powell, S., et al., eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Research, 2012. **40**: p. D284-D289.
131. Medini, D., et al., The microbial pan-genome. Current Opinion in Genetics & Development, 2005. **15**(6): p. 589-594.
132. Tettelin, H., et al., Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(39): p. 13950-13955.
133. Lefébure, T. and M.J. Stanhope, Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biology, 2007. **8**(5): p. R71.
134. Lukjancenko, O., T. Wassenaar, and D. Ussery, Comparison of 61 Sequenced *Escherichia coli* Genomes. Microbial Ecology, 2010. **60**(4): p. 708-720.
135. Skennerton, C.T., et al., Expanding our view of genomic diversity in *Candidatus* Accumulibacter clades. Environmental Microbiology, 2015. **17**(5): p.1574-1585.
136. Verberkmoes, N.C., et al., Shotgun metaproteomics of the human distal gut microbiota. ISME J, 2008. **3**(2): p. 179-189.
137. Karlin, S. and J. Mrázek, Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes. Journal of Bacteriology, 2000. **182**(18): p. 5238-5250.
138. VerBerkmoes, N.C., et al., Systems Biology: Functional analysis of natural microbial consortia using community proteomics. Nature Reviews Microbiology, 2009. **7**(3): p. 196-205.
139. Patti, G.J., O. Yanes, and G. Siuzdak, Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol, 2012. **13**(4): p. 263-269.

140. Frias-Lopez, J., et al., Microbial community gene expression in ocean surface waters. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(10): p. 3805-3810.
141. Wade, J.T. and D.C. Grainger, Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nature Reviews Microbiology, 2014. **12**(9): p. 647-653.
142. Régnier, P. and C.M. Arraiano, Degradation of mRNA in bacteria: emergence of ubiquitous features. BioEssays, 2000. **22**(3): p. 235-244.
143. Bernstein, J.A., et al., Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(15): p. 9697-9702.
144. Belasco, J.G., All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay. Nature Reviews Molecular Cell Biology, 2010. **11**(7): p. 467-478.
145. Chomczynski, P. and N. Sacchi, The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. Nature Protocols, 2006. **1**(2): p. 581-585.
146. Kopylova, E., L. Noe, and H. Touzet, SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics, 2012. **28**(24): p. 3211-7.
147. Stewart, F.J., E.A. Ottesen, and E.F. DeLong, Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. ISME J, 2010. **4**(7): p. 896-907.
148. He, S., et al., Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nature Methods, 2010. **7**(10): p. 807-12.
149. Lee, J.-H., H. Yi, and J. Chun, rRNASelector: A computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. The Journal of Microbiology, 2011. **49**(4): p. 689-691.
150. Garber, M., et al., Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Meth, 2011. **8**(6): p. 469-477.
151. Li, H. and R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 2009. **25**(14): p. 1754-1760.
152. Langmead, B. and S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nature Meth, 2012. **9**(4): p. 357-359.
153. Luo, R., et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 2012. **1**(1): p. 18.

154. Kearse, M., et al., Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 2012. **28**(12): p. 1647-1649.
155. Milne, I., et al., Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics, 2013. **14**(2): p. 193-202.
156. Dillies, M.-A., et al., A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics, 2013. **14**(6): p. 671-683.

Chapter 2: An Expanded Genomic Representation of the Phylum Cyanobacteria

Rochelle M. Soo[1], Connor T. Skennerton[1,2], Yuji Sekiguchi[3], Michael Imelfort[1], Samuel J. Paech[1], Paul G. Dennis[1†], Jason A. Steen[1], Donovan H. Parks[1], Gene W. Tyson[1,2], and Philip Hugenholtz[1,4*]

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia. [2]Advanced Water Management Centre, The University of Queensland, St Lucia, QLD 4072, Australia. [3]Biomedical Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), AIST Tsukuba Central 6, Ibaraki 305-8566, Japan[4]. Institute for Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia.

2.1 Abstract

Molecular surveys of aphotic habitats have indicated the presence of major uncultured lineages phylogenetically classified as members of the Cyanobacteria. One of these lineages has recently been proposed as a non-photosynthetic sister phylum to the Cyanobacteria, the Melainabacteria, based on recovery of population genomes from human gut and groundwater samples. Here, we expand the phylogenomic representation of the Melainabacteria through sequencing of six diverse population genomes from gut and bioreactor samples supporting the inference that this lineage is non-photosynthetic, but not the assertion that they are strictly fermentative. We propose that the Melainabacteria is a class within the phylogenetically defined Cyanobacteria based on robust monophyly and shared ancestral traits with photosynthetic representatives. Our findings are consistent with theories that photosynthesis occurred late in the Cyanobacteria and involved extensive lateral gene transfer, and extends the recognised functionality of members of this phylum.

2.2 Introduction

Cyanobacteria are recognised primarily for oxygenic photosynthesis [1], a feature that is thought to be common to all members of this phylum. However, oxygenic photosynthesis is widely thought to have originated after anoxygenic photosynthesis, likely well after the primary diversification of bacterial phyla [2]. This suggests that the Cyanobacteria as a primary bacterial line of descent predate oxygenic photosynthesis. Consistent with this inference is the relatively shallow

phylogenetic depth circumscribed by photosynthetic cyanobacteria compared to other bacterial phyla based on comparative rRNA analyses [3].

Cultured Cyanobacteria are categorised into five subsections on phenotypic grounds according to the botanical code [4]. However, over the past decade, 16S rRNA-based culture-independent molecular surveys have greatly increased our awareness of the phylogenetic breadth of the Cyanobacteria with the identification of additional major lines of descent such as YS2/4C0d-2, mle1-12, SM2F09, SM1D11 and ML635J-21 [5, 6]. These deep-branching Cyanobacteria have been found in numerous environments, including drinking water [7], grassland soil [8], wastewater treatment plants [9], and human and animal guts [10]. Many of these habitats are aphotic, which suggests that a large number of organisms phylogenetically defined as Cyanobacteria are non-photosynthetic [5, 6].

Recently, Di Rienzi *et. al.*, obtained five near complete genomes of members of the YS2/4C0d-2 lineage and confirmed the absence of photosynthetic machinery in these representatives. Additionally, comparative analysis of a 16S rRNA gene sequence (genbank acc. HM038000) obtained from freeze-dried cells of *Vampirovibrio chlorellavorus* ATCC 29753 [11] indicates that this bacterium is a member of the SM1D11 lineage. The original description of this organism provided no indication of photosynthetic capability further supporting absence of photosynthesis in basal cyanobacterial lineages. Di Rienzi and colleagues proposed a new phylum, Melainabacteria (Greek “nymph of dark waters”), for YS2/4C0d-2 and related basal lineages, given their deeply branching position relative to photosynthetic cyanobacteria and due to their lack of photosynthetic genes [12]. To further explore the inferred properties of the Melainabacteria group and to assess whether they should be excluded from the cyanobacterial phylum, we obtained six near-complete genomes representing a broader phylogenetic coverage of the Melainabacteria (representing YS2/4C0d-2, mle1-12 and SM2F09). Comparative analyses of these genomes corroborate the assertion that the Melainabacteria is a non-photosynthetic lineage, however they are robustly monophyletic with the photosynthetic cyanobacteria with which they share inferred common ancestral traits, such as cell envelope structure. We therefore suggest that the Melainabacteria represent a class within the Cyanobacteria extending the recognised metabolic capacity of this phylum.

2.3 Materials and Methods

A schematic of the workflow used in this study is presented in **Figure S2.1** with details as follows:

2.3.1 Sample collection and DNA extraction

Faecal samples from a 12 year old make koala (*Phascolarctos cereus*) named Zagget was collected in sterile 50ml falcon tubes on 12 May, 2011 (Zag_T1), 28 July, 2011 (Zag_T2) and 24 November, 2011 (Zag_T3), at Lone Pine Koala Sanctuary, Brisbane, Australia. Ethics approval for the collection of koala faeces was obtained from the Animal Welfare Unit, The University of Queensland, under ANRFA/074/11 *A study into Koala hindgut microbiology.* Samples were snap frozen in dry ice mixed with ethanol at the time of sampling and then transferred to -80°C until further processing. Genomic DNA was extracted from faeces using a MP-BIO FASTSPIN® spin kit for soil (MP Biomedicals, Solon, OH) according to manufacturer's instructions with the exception of two extra ethanol washes (**Table S2.1**).

Activated sludge was sampled from two, 4 L enhanced biological phosphorus removal (EBPR) sequencing batch reactors seeded from Thornside Wastewater Treatment Plant, Queensland, on 16 February, 2011. The first reactor (EBPR1) was operated on 6h reaction cycles of 120 min anaerobic phase, 180 min aerobic phase, 20 min settling, 20 min decant (1 L volume removed from the reactor supernatant) and 14 min pre-feed oxygen purge. At the end of each cycle, 1 L of nutrient solution containing 800 mg/L acetate and 40 mg/L phosphate (20:1 COD/P) was added over a period of 6 min (described in detail in [13]). The second reactor (EBPR2) was operated under identical conditions with the exception that the anaerobic phase lasted 60 min and nutrient solution was added over a period of 60 min compared to 6 min for EBPR1. Mixed liquor was collected 90 min into the aerobic phase and microbial biomass was concentrated by centrifugation at 4000 rpm for 2 min. Samples were collected at six timepoints from EBPR1 (EBPR1_T1 – EBPR1_T6) and at three timepoints from EBPR2 (EBPR2_T1 – EBPR2_T3) (**Table S2.1**). DNA was extracted from ca. 500 mg (wet-weight) of biomass with the MP-BIO FASTSPIN® spin kit for soil according to the manufacturer's instructions (MP Biomedicals, Solon, OH).

Methanogenic sludge samples were taken from a full-scale upflow anaerobic sludge blanket (UASB) reactor treating a high-strength organic wastewater discharged from a food-processing factory (Yamada et al. 2011). Two samples of the UASB sludge (A1 and A2) were taken on different dates (A1, 25, December, 2012; A2, 16 September, 2010). The A1 sample was further subsampled into two parts (flocculant sludge [F1] and granular sludge [G1]) by gravimetric settlement. Four samples in total (A1, A2, G1, F1) were used for sequencing. DNA was extracted by a bead-beating method as described previously [13](**Table S2.1**).

2.3.2 Community profiling of koala faeces and EBPR samples

The V6 to V8 regions of the 16S rRNA gene was amplified using fusion primers containing 454 adaptor sequences ligated to the primers 926F (5'-AAACTYAAAKGAATTGRCGG-3') and 1392R (5'-ACGGGCGGTGTGTRC-3') [14]. Multiplex identifiers consisting of five nucleotides were incorporated in the 1392R primer to allow for multiplexing. Fifty-microlitre PCR reactions were prepared containing 20 ng of template DNA, 5 μL of 10x buffer (Fisher Biotec, Wembley, Australia), 1 μL of 10 mM dNTP mix (Fisher Biotec) 1.5 μL BSAI (Fisher Biotech), 4 μL 25 mM $MgCl_2$ (Fisher Biotec), 1 μL of each 10 μM primer, and 1 unit of *Taq* polymerase (Fisher Biotec). Cycling conditions were 95°C for 3 min, followed by 30 cycles of 95°C for 30 s, 55°C for 30 s and 74°C for 30 s followed by a final extension of 74°C for 10 min. Following amplification, PCR products for each sample were purified using the Agencourt AMPure XP PCR purification system (Beckman-Coultier) and quantified using the Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA). Amplicons were sequenced from the reverse primers using the Roche 454 GS-FLX Titanium platform at the Australian Centre for Ecogenomics, University of Queensland, Australia (ACE, UQ). Sequence data generated were demultiplexed and processed using a modified version of the QIIME pipeline [15], which uses Acacia 1.50 (app v 2.0.0) [16] to correct homopolymer errors (modified pipeline is available at https://github.com/Ecogenomics/APP). Sequences were clustered at 97% sequence identity and the taxonomy of the representatives from each OTU was assigned using blastn v. 2.2.26 [17] against the Greengenes database, version 12_10) [18].

2.3.3 Community profiling of UASB samples

For microbial community structure analysis of the UASB samples, the V4 regions of the 16S rRNA gene were amplified using fusion primers for the Illumina sequencing platform [19]. Multiplex identifiers of twelve nucleotides were incorporated in the M806R primer to allow for multiplexing. PCR conditions were described elsewhere [13]. Following amplification, PCR products for each sample were purified using the Agencourt AMPure XP PCR purification system (Beckman-Coultier) and quantified using the Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA). Amplicons were sequenced from both the forward and reverse primers using the Illumina MiSeq platform and the MiSeq 500 cycles reagent kit v2 (Illumina Inc.). Sequence data generated were demultiplexed and processed using the QIIME pipeline [15]. Sequences were clustered at 97% sequence identity and the taxonomy of the representatives from each OTU was assigned using blastn [20] against the Greengenes database [18].

2.3.4 Paired end sequencing

The genomic DNA from Zag_T2, EBPR1_T1, EBPR1_T3, EBPR1_T5, EBPR2_T1, EBPR2_T2, EBPR2_T3 and were sent to Aalborg University, Denmark where DNA libraries were prepared for sequencing using TruSeq DNA Sample Preparation Kits v2 (Illumina, San Diego, CA, USA) with 2 μg of DNA following the manufacturer's instructions with nebuliser fragmentation. Library DNA concentration was measured using the QuantIT kit (Molecular probes, Carlsbad, CA, USA) and paired-end sequenced (2 x 150 bp with an average 250 bp fragment size) on an Illumina HiSeq2000 using the TruSeq PE Cluster Kit v3-cBot-HS and (Illumina). The Zag_T2 sample was sequenced on a whole lane of a flowcell and the EBPR1_T1, EBPR1_T3 and EBPR1_T5 were sequenced on a third of a flowcell lane each.

Zag_T1 and Zag_T3, EBPR1_T2, EBPR1_T4, EBPR1_T6, were sequenced at the Institute for Molecular Bioscience, The University of Queensland (IMB, UQ) generating paired-end 150 bp reads (with an average fragment size of 320) using the Nextera DNA Sample Prep kit (Illumina) and the Illumina HiSeq2000 platform. Each Zag sample was sequenced on a quarter of a flowcell lane each and the EBPR samples was sequenced on a third of a flowcell lane each (**Table S2.1**).

DNA extracts from the four UASB sludge samples (A1, A2, F1, F2) were fragmented to 250-400bp using a Covaris S2 (Covaris, Woburn, MA, USA), and were used for library preparation with a TruSeq sequencing kit (Illumina). Library DNA concentration was measured using the QuantIT kit (Molecular probes) and paired-end sequenced (2 x 250 bp with an approximate average fragment size of 300 bp) on an Illumina MiSeq system using the MiSeq 500 cycles reagent kit v2 (Illumina Inc.). Each of the four DNA libraries was sequenced in a single MiSeq run (**Table S2.1**).

Raw paired end 2 x 75 bp Illumina data for two male and five female gut metagenome datasets were downloaded from the public MetaHIT database. Details relating to the collection, sequencing, and analysis of the MetaHIT data are provided at http://www.metahit.eu/ [21] (**Table S2.1**).

2.3.5 Sequence assembly and population genome binning

Paired end reads for the koala faeces, MetaHIT, EBPR, and UASB samples were quality trimmed using CLC workbench v6 (CLC Bio, Taipei, Taiwan) with a quality score threshold of 0.01 (phred score 20) and minimum read lengths as follows; 100 bp for the koala faecal and EBPR and samples, 125 bp for the UASB samples, and 50 bp for the MetaHIT samples, in accordance with the read length for each dataset. No ambiguous nucleotides were accepted and Illumina sequencing adapters were trimmed if found.

Trimmed sequences for each biome were assembled using CLC's *de novo* assembly algorithm, with a kmer size of 63 for koala faecal, EBPR, and UASB samples and a kmer size of 45 for the MetaHIT data.

Population genomes were recovered from the paired-end assemblies using GroopM, version 1.0 with default settings [22]. Briefly, reads from each sample were mapped onto their corresponding co-assemblies (scaffolds ≥ 500 bp; **Table S2.1**) and coverage patterns for each scaffold were calculated, transformed and projected onto a 3-dimensional plot in which scaffolds from the same population genome would cluster. Integrity of bins was initially confirmed using the GroopM visualisation tool.

2.3.6 Population genome completeness and contamination

All contigs in each population genome bin were translated into six open reading frames and a set of 105 single copy marker genes (a subset of the 111 single copy marker genes widely conserved in Bacteria from [23]) were identified in the translated dataset using HMMER3 [24] with default settings and the model-specific PFAM [25] and TIGRFAM [26] thresholds. Completeness was estimated as the percentage of the 105 markers identified in any given population bin, and contamination as the percentage of markers found in >1 copy in a population bin (**Table 2.1**). The marker gene identification and completeness/contamination calculation functions are combined in the software tool CheckM version 0.5.0 [27].

2.3.7 Taxonomic assignment of population genomes

To identify putative representatives of the Melainabacteria amongst the population bins of a minimum quality threshold (>60% completeness, <10% contamination), we constructed a maximum likelihood tree based on a concatenated set of 83 marker genes with known reference genomes (see *Whole genome phylogeny* below) including previously reported standard draft Melainabacteria genomes [12].

2.3.8 Mate pair sequencing for Melainabacteria genome improvement

Genomic DNA extracted from Zag_T1, Zag_T2 and Zag_T3 were multiple strand-displacement amplified in triplicates using the Illustra Genomiphi V2 DNA amplification kit (GE Healthcare) as per manufacturer's instructions. 1 μg of DNA was used for Mate-pair libraries using the Illumina MiSeq sequencing protocol and a gel-free protocol (2-15 kbp inserts).

EBPR1_T1, EBPR1_T6, EBPR2_T1 and EBPR2_T3 were sequenced using long-insert mate-pair sequencing according to Illumina's Nextera protocol. DNA was size selected for 3.5 kbp fragments with a standard deviation of 300bp and further sequenced at IMB, UQ with the Illumina HiSeq platform. Raw mate-pair reads were processed using Prepmate 0.2, removing read pairs where less than 30 bp remained after trimming the adaptor sequence (https://github.com/ctSkennerton/prepmate). Approximately 50% of the raw reads were retained as valid mate-pairs (reads correctly oriented in the reverse-forward direction) resulting in between 16 to 19 million read pairs per sample (**Table S2.1**).

For the UASB samples, one out of the four samples, A1, was sequenced at the National Institute of Advanced Industrial Science and Technology, Japan (AIST) using the Mate Pair Library Preparation Kit v2 (Illumina) and the MiSeq 500 cycles reagent kit v2 (Illumina) on an Illumina MiSeq system.

Mate-pair sequence data for each sample type (except the public MetaHIT data) was used to scaffold contigs in identified Melainabacteria population genome bins (**Table S2.1**) using the default settings of SSPACE v2.0 [28]. Scaffolded assemblies were then checked from completeness and contamination using CheckM (**Table 2.1**).

2.3.9 16S rRNA gene reconstruction

16S rRNA genes are often difficult to recover via differential coverage binning due to coassembly of rRNA genes present in multiple copies, which distorts their coverage statistics. Therefore, 16S rRNA genes were independently reconstructed from the metagenomic data by extracting read pairs that matched an HMM model of the 16S rRNA gene built using HMMER v3.1b1 from 42,363 bacterial and 1,119 archaeal sequences within the 94% dereplicated Greengenes database released on October, 2012 [5]. These extracted read pairs were then mapped to the Greengenes database using BWA-MEM v0.7.5a-r405 [29]. A read was considered reliably mapped if at least 85% of the read aligned to the reference sequence, and the edit distance of the alignment was at most 10% of the length of the read (e.g., less than 10 for 100 bp reads. Pairs were further filtered to remove any pair where both reads did not properly map to reference sequences within a branch length of 0.03 as measured over the Greengenes phylogeny.

The remaining pairs were clustered in a greedy manner in order to identify pairs mapping to similar Greengenes reference sequences. Reference sequences were put in ascending order according to the number of pairs mapped to them. Starting with the reference sequence with the highest number of

assigned pairs, any pairs assigned to a reference sequence within a branch length of 0.03 to this reference sequence were clustered together and removed from further consideration. This process was repeated until all pairs were assigned to a cluster. Each cluster of 16S pairs was then independently assembled using CLC Workbench v6.5. Using this technique an additional 16S rRNA gene was recovered from Zag_221 (1,403 bp).

2.3.10 16S rRNA phylogeny

16S rRNA genes from Melainabacteria population genomes were aligned to the standard Greengenes alignment with PyNAST [5]. Aligned sequences and a Greengenes reference alignment, version gg_13_5 (ftp://greengenes.microbio.me/ greengenes_release; [5]) were imported into ARB [30] and the Melainabacteria sequence alignments were manually corrected using the ARB EDIT tool. For constructing the alignment data of different taxon configurations, representative taxa (>1,300 nt) were selected and their alignment data were exported from ARB with the Lane mask filtering, resulting in totals of 402 and 67 taxa for two data sets. Neighbour joining trees were calculated from the masked alignments with LogDet distance estimation using PAUP*4.0 [31]. A further analysis was run with 100 bootstrap replicates. Maximum parsimony trees were calculated using PAUP*4.0 [31]. A heuristic search was used with a random stepwise addition sequence of 10 replicates and nearest-neighbour-interchange swapping. A further analysis was run with 100 bootstrap replicates. Maximum likelihood trees were calculated from the masked alignments using the Generalised Time-Reversible model with Gamma and I options in RAxML version 7.7.8 [32] (raxmlHPC-PTHREADS -f a -k -x 12345 -p 12345 -N 100 -T 4 -m GTRGAMMAI). Bootstrap resampling data (100 replicates) were generated with SEQBOOT in the phylip package [33], and were used for 100 bootstrap resamplings. Generated trees were re-imported into ARB for visualisation (**Figure 2.1B**). A BLASTN analysis was performed on the 16S rRNA genes extracted from the koala bins and the 16S rRNA gene from *Eucalyptus grandis* chloroplasts to check for chloroplast contamination.

2.3.11 Whole genome phylogeny

Two sets of ubiquitous single copy marker genes were obtained and aligned from seven high-quality and four standard draft Melainabacteria population genomes (**Table 2.1**) and up to 434 complete bacterial and archaeal reference genomes obtained from IMG (v4.0) [34] using HMMER3 [24]. The first set of 38 markers [32] is found in nearly all Bacteria and Archaea and the second set of 83 markers (**Table S2.2**) was derived from a subset of the 111 bacterial marker set [23] based on congruent taxonomic signal as follows. Individual gene trees were constructed using FastTree v2.1.7 [35] and compared to the IMG taxonomy [34]. We used the following measure to quantify

the agreement of each node in an unrooted gene tree with a specific clade *c* (e.g., Bacteria, Firmicutes) within the IMG taxonomy:

$$\text{consistency} = \max(N_L(c) / (T(c) + I_L(c)) , N_R(c) / (T(c) + I_R(c))$$

where T(*c*) is the total number of genomes from clade *c*, the subscripts *R* and *L* indicate the subset of genomes to the 'right' and 'left' of the node under consideration, $N_x(c)$ is the number of genomes in subset *x* from clade *c*, and $I_x(c)$ is the number of genomes in subset *x* not from clade *c*. The consistency of a clade *c* was assigned the highest consistency found over all nodes. Average consistencies were then determined over all clades with at least five genomes independently at the domain, phylum, and class ranks. Gene trees where the average consistency over these three ranks was less than 0.86 were discarded as a sharp drop-off in consistency was observed beyond this threshold.

Ambiguous and uninformative alignment positions were removed from aligned sets of concatenated marker genes using gblocks [36] under default settings with the exception that a conserved position was not allowed to have gaps in more than half of the sequences. Phylogenetic trees were reconstructed from the two filtered marker gene alignments with outgroup configurations as detailed in Tables S3 and S4**.** All tree topologies were tested for robustness using the maximum likelihood methods from FastTree version 2.1.7 (JTT model, CAT approximation) [35], RAxML version 7.7.8 with a JTT and Gamma models [37] (raxmlHPC-PTHREADS -f a -k -x 12345 -p 12345 -N 100 -T 8 -m PROTGAMMAJTT), and maximum parsimony method using PAUP*4.0 with heuristic search, a random stepwise addition sequence of 10 replicates, and nearest-neighbour-interchange swapping. Generated trees were imported into ARB where they were rooted, beautified and grouped for display purposes.

2.3.12 Melainabacteria genome annotation and metabolic reconstruction

The draft Melainabacteria genomes were submitted to IMG/ER [38] for automated annotation and manual analysis. KEGG maps and gene annotations were used to reconstruct the metabolism of the Melainabacteria representatives and a composite metabolic cartoon was prepared in Adobe Illustrator CS6 (**Figure 2.2** and **Table S2.5**).

Average nucleotide identity was calculated using the ANI calculator with default settings (http://enve-omics.ce.gatech.edu/ani/).

2.3.13 Protein family analysis

The presence of Pfams and TIGRfams for maximally differentiating cell wall types, as previously described in [39] and flagella assembly [40], were identified in the draft Melainabacteria genomes and 2,363 representative phyla using complete bacterial genomes obtained from IMG (v4.0) [34] (**Table S2.6**). Photosynthesis and (bacterio)chlorophyll biosynthesis genes as described in [41] (**Table S2.6**) were also identified as present or absent by using the BLASTP module [42] in IMG with an e-value of >1e-10 and amino acid identities of ≥ 25%. Paralogs from the cobalamin pathway or later steps in the bacteriochlorophyll c pathway were removed. The colour key represents the number of species within a certain phylum that have the Pfams, TIGRfams or genes versus the total number of complete bacterial genomes obtained from IMG or the draft Melainabacteria genomes. A heat map was constructed in RStudio v0.95.265 [43] using gplots [44] and RColorBrewer [45] (**Figure 2.3**).

COG profiles for each genome were constructed through homology search between ORFs predicted with Prodigal v2.60 [46] and the 2003 COG database [47]. Homology was assessed with BLASTP v2.2.26+ [20] using an e-value threshold of 1e-5 and a percent identity threshold of 30%. The relative percentage of each COG was calculated in relation to the total number of ORFs predicted for each genome. All statistical plots and analyses were conducted using STAMP v2.0.1 [48].

2.4 Results and Discussion

During ongoing culture-independent molecular surveys, 16S rRNA phylotypes belonging to basal cyanobacterial lineages were identified in a number of habitats. These included faecal samples collected from a geriatric male koala (*Phascolarctos cinereus*; Zagget), a lab-scale sequencing batch reactor performing enhanced biological phosphorous removal (EBPR) and an upflow anaerobic sludge blanket (UASB) reactor treating a high-strength organic wastewater discharged from a food processing factory (see sampling details below). In parallel, public metagenomic datasets of human faecal samples containing members of YS2/4C0d-2 [21] were re-analysed with the goal of obtaining additional genomes from this lineage.

Three distantly related (88% 16S rRNA gene identity) phylotypes belonging to the YS2/4C0d-2 lineage were detected in the koala faeces, a representative of mle1-12 was identified in the EBPR bioreactor and a representative of SM2F09 was identified in the UASB reactor (**Figure 2.1**). Although Melainabacteria are typically found in low abundance, these phylotypes comprised up to 6.7%, 4.2% and 1.7% of the koala faecal, EBPR and UASB microbial communities respectively. Samples with the highest relative abundance of Melainabacteria were chosen for deep metagenomic

sequencing to improve the likelihood of obtaining near-complete population genomes for comparative analysis. The relative abundance of the Melainabacteria in the MetaHIT shotgun datasets was estimated to be up to 2.6% by direct classification of 16S rRNA genes in the shotgun datasets.

2.4.1 Recovery of Melainabacteria population genomes

Koala faecal samples were collected from three timepoints from the same koala over a period of six months and sequenced to produce 90.7 Gbp of metagenomic data. Two EBPR reactors were sampled six times and three times respectively over a period of seven months, producing 211.6 Gbp, and the UASB reactor was sampled twice producing 31.5 Gbp of metagenomic data (**Table S2.1**). Human faecal metagenomes from healthy Danish individuals (two male and five female) were obtained from the public repository (http://www.ebi.ac.uk/ena/home, study *accession number ERP000108*) comprising a total of 21.6 Gbp. Multiple datasets from the same sample types were co-assembled which produced between 3,139 and 148,338 contigs with an N50 of 1.4 kbp to 4.6 kbp. Population genomes were extracted from the assemblies using differential coverage binning of each set of related metagenomes. This approach leverages differences in relative abundance of target populations between two or more related samples to identify contigs with co-varying coverage [39, 49]. Population genomes obtained using this method were taxonomically assigned by placement in concatenated gene trees comprising all finished IMG reference genomes [34] (**Table S2.2;** *see below*). Six population genomes were found to form a monophyletic lineage together with the reference cyanobacterial genomes (**Figure 2.1A**). These comprised one mle1-12 representative from the EBPR (EBPR_351), one SM2F09 representative from the UASB reactor (UASB_169), and four YS2/4C0d-2 representatives from koala and human faeces (Zag_1, Zag_111, Zag_221 and MH_37) consistent with 16S rRNA gene amplicon community profiling. Analysis of 16S rRNA genes recovered from four of the six population genomes also confirmed that they are members of the Melainabacteria (**Table 2.1** and **Figure 2.1B**) and not chloroplast contamination (79% identity with the 16S rRNA gene from *Eucalyptus grandis* chloroplast). Further sequencing using long-insert (2-15 kb) libraries of the EBPR, UASB reactor and koala faeces were used to improve the quality of the draft population genomes (**Table S2.1**). The completeness and degree of contamination of the improved genomes was estimated by determining the presence and number of copies of 105 conserved single-copy bacterial marker genes [23]. All population genomes had >90% estimated completeness (>97 of the 105 conserved single-copy bacterial marker genes) and <10% contamination (multiple copies of genes expected to be present as single copies) (**Table 2.1**).

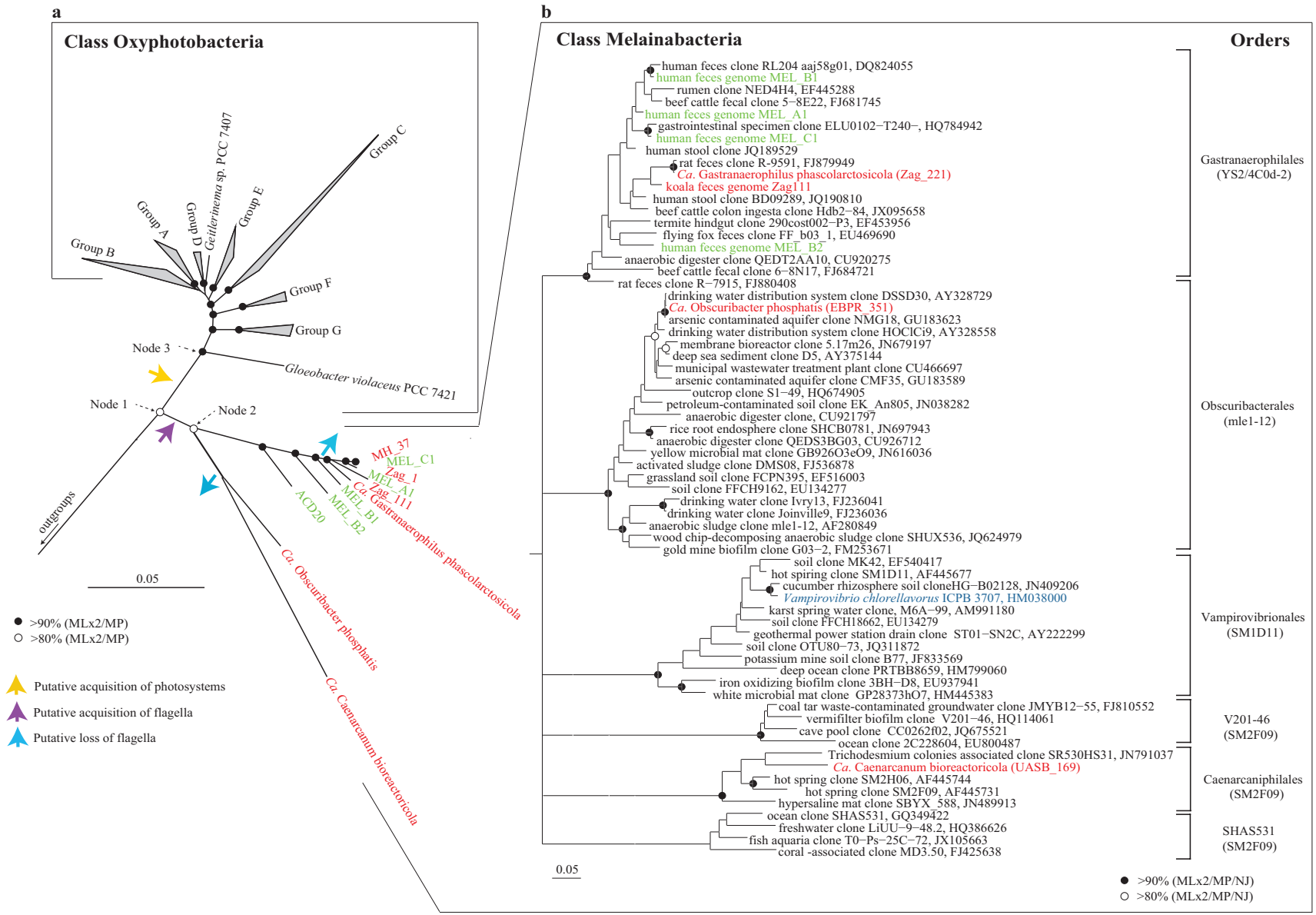
2.4.2 An expanded phylogenetic classification of the phylum Cyanobacteria

We began analysis of the Melainabacteria genomes by constructing phylogenetic trees based on two concatenated alignments of broadly conserved single copy marker genes (*see Methods*; **Table S2.2** and [32]). The ingroup comprised 81 reference cyanobacterial genomes and 11 Melainabacteria genomes; six determined in the present study and the five most complete genomes obtained in the Di Rienzi *et. al.* (2013) study (**Figure 2.1A** and **Table S2.3**). We evaluated the monophyly of the photosynthetic Cyanobacteria and Melainabacteria groups using up to 377 outgroup genomes representing 28 phyla (**Figures S2.2 to S2.4** and **Table S2.4**). In all cases, the evolutionary association between the two groups was reproducibly resolved with >80% bootstrap support (node 1, **Figure 2.1A**). Di Rienzi *et. al.* (2013) concluded that the two groups are sister phyla rather than a single phylum based on a combined divergence slightly greater than a recommended threshold of 85% 16S rRNA gene sequence identity for distinguishing new phyla [50]. However, the primary criterion for defining a new phylum is not satisfied in this instance, i.e. that the lineage is reproducibly unaffiliated with existing phyla [50] according to both 16S rRNA and genome-level phylogenies. There are well known examples of reproducibly associated (sister) phyla in the bacterial domain grouped into superphyla, such as the PVC (Planctomycetes, Verrucomicrobia, Chlamydiae) superphylum [52], that arguably should be consolidated into single phyla according to the Hugenholtz *et al.* (1998) definition. We suggest that the PVC and other superphyla are historical artifacts that should be consolidated into phyla to provide a more naturalistic taxonomy based on reproducible divergence points in evolutionary trees.

Several inferred features beyond the evolutionarily conserved core set of genes used to construct the genome tree are consistent with a common ancestry between the photosynthetic cyanobacteria and Melainabacteria (see below). Therefore, we propose that the phylum Melainabacteria should be reclassified as the class Melainabacteria within the phylum Cyanobacteria. Me.lai.na.bac.te.ria. Gr. n. Melaina, a nymph in Greek mythology, who presides over dark subterranean aspects; N.L. masc. n. *bacter* (from Gr. n. *baktron*), a rod; suff -*ia* ending to denote a class; N.L. fem. pl. n. *Melainabacteria* class of bacteria found in the dark.

The node defining the Melainabacteria in the concatenated gene alignment tree (node 2, **Figure 2.1A**) was supported in all analyses with >80% confidence consistent with the genome-based analysis of Di Rienzi *et. al.* (2013). The population genomes formed three primary lines of descent within the Melainabacteria, with the human and koala gut genomes and groundwater genome ACD20 [12] forming a monophyletic cluster. 16S rRNA-based inference provides only modest

Figure 2.1. Concatenated gene tree of the phylum Cyanobacteria and 16S rRNA gene tree of class Melainabacteria



(a) A maximum likelihood phylogenetic tree based on the concatenated alignment of 83 phylogenetically conserved proteins (**Table S2.2**) is shown as a concatenated gene tree of the phylum Cyanobacteria, where Chloroflexi genomes are used as the outgroup (**Table S2.3**). Groups A to G for the Oxyphotobacteria were based on the names given by [51]. *Candidatus* has been abbreviated to *Ca.* Bootstrap resampling analyses (100 times) with maximum likelihood (ML; using RAxML and FastTree) and maximum parsimony (MP; using PAUP*) methods were performed for different taxon configurations with 38 or 83 conserved proteins (**Figures S2.1 to S2.3**), and the phylogenetic robustness (monophyly score) of taxa is indicated at the node: Black circles in the tree represents nodes with >90% bootstrap supports by all calculations, white circles represents nodes with >80% bootstrap supports by all calculations. The acquisition of photosystems is proposed to have occurred between node 1 and node 3, whereas the acquisition of flagella occurred between node 1 and node 2 and was subsequently lost at two time points. Genomes in green are representatives from Di Rienzi *et al.,* 2013 and genomes in red are Melainabacteria from this study. **(b)** A maximum likelihood phylogenetic tree based on 16S rRNA genes from the class Melainabacteria recovered from Di Rienzi *et al.,* 2013 and this study, together with public representatives from the Greengenes database version 13_05 and Silva 115 database. 16S representatives in green are Melainabacteria representatives from Di Rienzi *et al.,* 2013, the Melainabacteria in red are representatives from this paper and the 16S from the cultured representative, *Vampirovibrio chlorellavorus,* is highlighted in blue. Bootstrap analyses with maximum likelihood (ML; using RAxML and FastTree), maximum parsimony (MP; using PAUP*), and neighbour joining (NJ; using PAUP*) methods were performed (100 times resampling), and the monophyly score is indicated at the node: black circles represents nodes with >90% bootstrap support with all calculations and white circles represents nodes with >80% bootstrap support. We propose that there are at least six orders in the class Melainabacteria, and a third class-level lineage in the Cyanobacteria, ML635J-21 (**Figure S2.4)**.

Table 2.1. Summary statistics for population genomes belonging to the class Melainabacteria

Population genome	# of scaffolds	Estimated genome size (Mbp)	%GC	Number of genes[a]	rRNAs[b]	Estimated completeness (%)[c]	Estimated contamination (%)[c]	Proposed candidatus name	Study
Zag_221	14	1.8	38.5	1838 (1799)	16S	100.0	1.0	Gastranaerophilus phascolarctosicola	Present study
Zag_1	322	2.0	34.9	2194 (2160)	-	94.3	1.9		Present study
Zag_111	65	2.2	36.7	2313 (2257)	5S, 16S 23S	98.1	5.7		Present study
MH_37	157	2.2	34.1	2402 (2360)	-	100.0	1.0		Present study
MEL_A1	1	1.9	33.0	1879 (1832)	5S, 16S, 23S	100.0	2.9		Di Rienzi et al., 2013
MEL_B1	21	2.3	35.4	2269 (2219)	5S, 16S, 23S	100.0	1.0		Di Rienzi et al., 2013
MEL_B2	26	2.3	36.3	2262 (2215)	5S, 16S, 23S	100.0	1.9		Di Rienzi et al., 2013
MEL_C1	4	2.1	34.1	2162 (2120)	5S, 16S, 23S	100.0	1.9		Di Rienzi et al., 2013
ACD20[d]	185	2.7	33.5	2455 (2325)	5S, 23S	100.0	2.9		Di Rienzi et al., 2013
EBPR_351	8	5.5	49.4	4392 (4342)	5S, 16S, 23S	99.1	7.6	Obscuribacter phosphatis	Present study
UASB_169	67	1.8	27.5	1917 (1870)	16S, 23S	94.3	0.0	Caenarcanum bioreactoricola	Present study

[a]Numbers in brackets for number of genes is the number of protein coding genes.

[b]16S rRNA lengths are >1,000 bp.

[c]Estimated completeness and estimated contamination is based on the 105 single copy marker genes (a subset of the 111 single copy marker set from Dupont *et al.,* 2012).

[d]ACD20 is the corrected genome from Albertsen *et al.,* (2013) as the original completeness for ACD20 was 100.0% and original contamination was 107.6%.

support for the monophyly of the Melainabacteria (**Figure 2.1B** and **Figure S2.5**) and indeed several of the robustly monophyletic groups therein are classified as primary cyanobacterial lines of descent (classes) in Greengenes and Silva [5, 6]. The higher resolution afforded by the genome sequences suggests that these lineages should be classified as orders within the class Melainabacteria (**Figure 2.1**). We propose names for four of these orders based on habitat and analysis of the population genomes (*see below*), and the recognition of *Vampirovibrio chlorellavorus* in the SM1D11 lineage (**Figure 2.1**). The location of the ACD20 genome could not be determined within the 16S rRNA gene tree (**Figure 2.1B**) as it lacks a 16S rRNA gene sequence, but genome trees based on a refined binning of this population [39] (**Table 2.1**) indicate that it is basal and monophyletic with the order, Gastranaerophilales (YS2/4C0d2; **Figure 2.1A** and **Figures S2.2** and **S2.3**). An additional distantly related cyanobacterial lineage ML635J-21 [5, 6] currently not represented by a sequenced genome may represent another class-level lineage within the Cyanobacteria (**Figure S2.4**) highlighting the need for further genomic exploration of the cyanobacterial phylum.

Our analyses show that the photosynthetic cyanobacteria are robustly monophyletic within the expanded context of the phylum Cyanobacteria (node 3, **Figure 2.1A**). We therefore propose to reinstigate the name Oxyphotobacteria [53] to describe all photosynthetic cyanobacteria (including chloroplasts) in a single class. Gr. adj. *oxus*, acid or sour and in combined words indicating oxygen; Gr. n. *phos photos*, light; Gr. n. *baktêria*, staff, cane; suff. -*ia*, ending to denote a class; N.L. neut. pl. n. *Oxyphotobacteria*, light-requiring bacteria that produce oxygen. The name implies that the class is able to photosynthesise. We denote the order-level groupings within this class as A to G (**Figure 2.1A**) in accordance with a recent genome-based analysis [51]. Oxyphotobacteria are still classified primarily on morphological grounds into five subsections [4] despite clear incongruencies between phylogenetic reconstructions and morphological complexity [51]. Therefore, it is likely that the order- and family-level groupings within the Oxyphotobacteria will be reclassified on phylogenetic grounds with a concomitant widespread reclassification of cyanobacterial strains once this group is no longer under the jurisdiction of the Botanical Code [54].

2.4.3 Inferred metabolism of Melainabacteria genomes

Di Rienzi *et, al*. inferred metabolic properties of the class Melainabacteria based on comparative analysis of draft population genomes belonging to only one order, the Gastranaerophilales (**Figure 2.1**). We substantially increase the phylogenetic coverage of the Melainabacteria in the present study by recovery of population genomes spanning three of the six identified orders (**Figure 2.1B**). Expanded genomic representation should provide a more balanced overview of the metabolic

properties of this class including features in common with, or distinct from, the Oxyphotobacteria. We begin by proposing Candidatus species for the most complete genomes in each of the three orders obtained in this study and describe their inferred metabolic properties below.

The most complete Gastranaerophilales genome with the least number of scaffolds, Zag_221, was selected as the Candidatus species representative of the group, *Candidatus* Gastranaerophilus phascolarctosicola (**Table 2.1** and **Figure 2.1**). *Gastranaerophilus phascolarctosicola* (Gas.tra.nae.ro.phi'lus. Gr. n. *gaster* stomach; Gr. pref. *an*-, not; Gr. masc. n. *aer*, air; L. masc. adj. *philus* [from Gr. adj.*philos*], friend, loving; *Gastranaerophilus* a bacterium loving anaerobic gastric environments. 'phas.co.larc.to.si.co.la'. N.L. *Phascolarctos* the name of koala; L. suffix *-cola* inhabitant, dweller; N.L. masc. n. *phascolarctosicola* hiding in the belly of a koala). The genome size and GC content range of the four Gastranaerophilales genomes were in accord with the Di Rienzi *et. al.* population genomes from this order (**Table 2.1**). Members of this group have small streamlined genomes ranging in size from 1.8 to 2.3 Mb, with the exception of ACD20, which is 2.7 Mb after binning refinement [39]. The Gastranaerophilales genomes recovered from the koala and human faeces in the present study support the assertion [12] that this lineage comprises obligate fermenters missing the genes necessary for aerobic and anaerobic respiration, as well as the tricarboxylic acid (TCA) cycle (**Figure 2.2** and **Table S2.5**). Instead, all Gastranaerophilales genomes contain the Embden-Meyerhof-Parnas (EMP) pathway, capable of converting glucose, mannose, starch or glycogen into lactate, ethanol and/or formate (**Figure 2.2**). All representative genomes have the potential to produce riboflavin, nicotinamide, biotin, dehydrofolate and pantoate as found previously [12].

Di Rienzi *et. al.* highlighted the presence of FeFe hydrogenases in their human gut population genomes speculating that these organisms are hydrogen-producing anaerobes in syntrophic interactions with hydrogenotrophic methanogens or acetogens. We also identified FeFe hydrogenases in the MH_37 genome obtained from the human gut, but in contrast found Fe-only or NiFe hydrogenases in koala gut Gastranaerophilales genomes (**Figure S2.6**). It is possible that the less oxygen sensitive NiFe hydrogenase would allow members of this order to colonize the jejunum as well as the more anaerobic colon [55].

Di Rienzi *et. al.* also reported that the Melainabacteria are flagellated based on the presence of a complete set of flagella genes in three of their draft Gastranaerophilales genomes (MEL_B1, MEL_B2 and ACD20). Of the nine Gastranaerophilales genomes available (**Table 2.1**), only these three had complete flagella gene sets, the remainder having only a subset of genes that would not

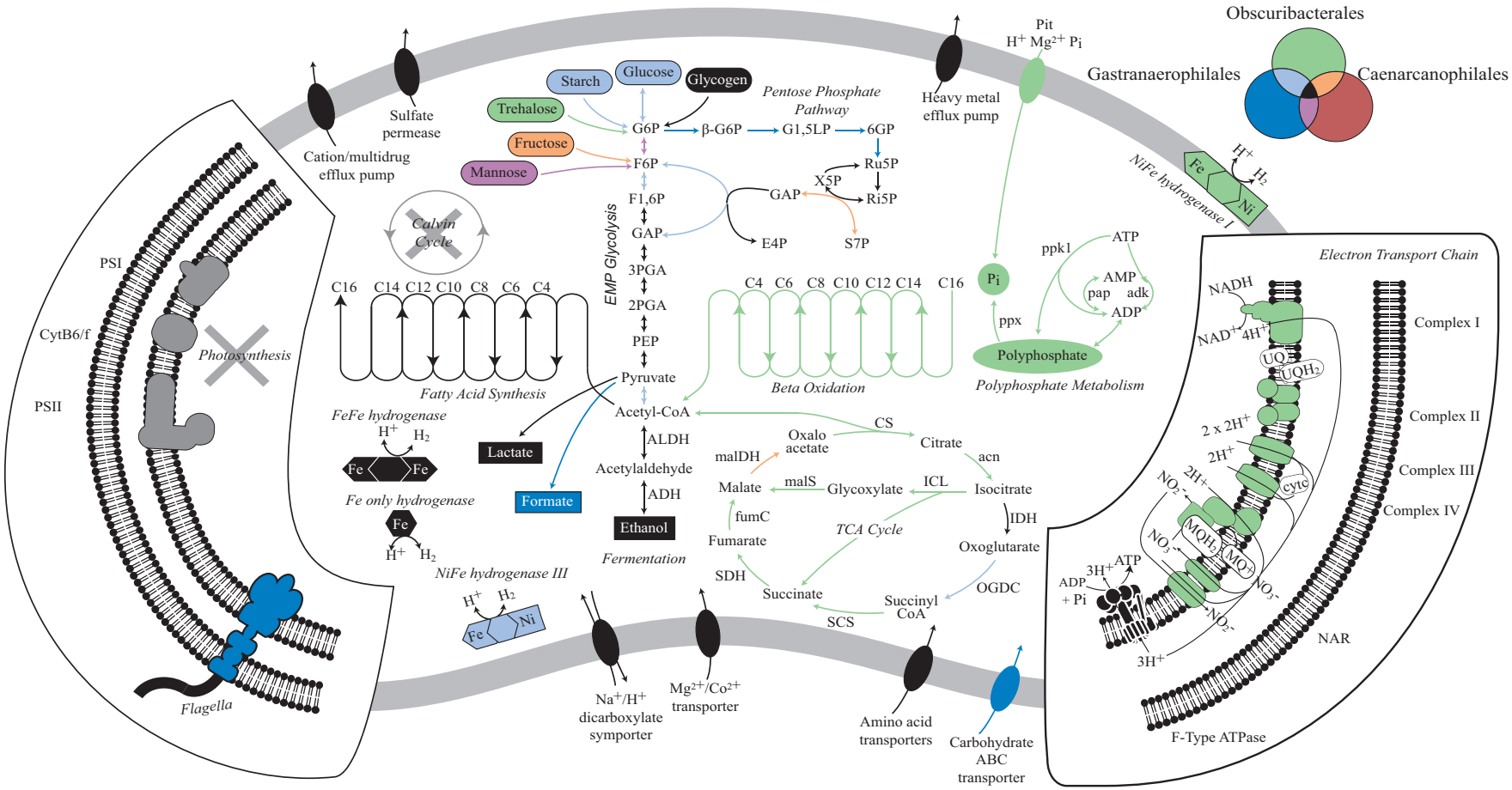
encode a functional flagellum (**Figure S2.7**). We infer that flagella were present in the ancestor of the class Melainabacteria and subsequently lost on at least two occasions based on monophyly of flagella genes common to the Gastranaerophilales and Caenarcaniphilales (**Figure 2.1B** and **Figure S2.7)**. Moreover, there appears to have been a subsequent loss of functional flagella in *'G. phascolarctocola'* and relatives (**Figure S2.7**) indicating the presence of non-motile members of this order in animal gut habitats.

Based on a suggested species threshold of 95% average nucleotide identity [56], three of the Gastranaerophilus genomes can be considered to belong to the same species (**Table 2.1** and **Figure S2.8**). Two of these genomes were recovered from humans (MH37 and MEL_C1) and the third from a koala (Zag_1) despite the structural and physiological differences between the human and koala gut. The majority of genes that differ between these genomes are hypothetical proteins or phage associated (**Figure S2.8**), typical of differences seen between strains belonging to the same species.

Candidatus Obscuribacter phosphatis (Ob.scur.i.bac.ter. L. adj. *obscurus* dark; N.L. masc. n. *bacter* (from Gr. n. *baktron*), a rod; *Obscuribacter* a bacterium found in the dark. 'phos.pha.tis'. N.L. n. *phosphatis*, phosphate; N.L. *phosphatis* accumulating phosphate) EBPR_351 (*'O. phosphatis'*) representing the order Obscuribacterales (**Figure 2.1** and **Table 2.1**) is conspicuous among the Melainabacteria genomes because of its larger size (5 Mb) and associated metabolic versatility. '*O. phosphatis'* contains the genes necessary for polyphosphate metabolism, including a low affinity inorganic phosphate transporter (PiT), polyphosphate kinase 1 (used to synthesise or degrade polyP while consuming or generating respectively ATP directly), polyphosphate kinase 2 (degrades polyP producing GTP from GDP), exopolyphosphatase (degrades polyP in a non-reversible reaction that does not generate ATP directly), polyphosphate:AMP phosphotransferase and adenylate kinase (Seviour & Nielsen 2010). *'O. phosphatis'* has the capacity for aerobic and anaerobic respiration, and fermentation, allowing it to function during both the oxic and anoxic phases of EBPR [57]. It contains genes encoding a complete respiratory chain including Complexes I, II, III and IV and an F-Type ATPase (**Figure 2.2**). Di Rienzi *et. al.* concluded that the Melainabacteria lack electron transport chains and are therefore incapable of respiration. This highlights the dangers of inferring phylum- or class-level functionality based on limited phylogenetic sampling of the lineage.

Like the Gastranaerophilales, *'O. phosphatis'* has the capability to metabolise a wide range of simple carbohydrates via the EMP pathway, and also fatty acids via the beta-oxidation pathway

Figure 2.2. Metabolic reconstruction of Melainabacteria representatives



Metabolic predictions for Gastranaerophilales, Obscuribacterales and Caenarcaniphilales representatives based on genome annotations. Names of pathways are italicised and fermentation products are shown in rectangles. Features identified in one or more of the orders are highlighted by colour. Missing pathways are shown in grey. The Melainabacteria are missing photosystems I and II, and the Calvin Cycle, which are found in Oxyphotobacteria. All Melainabacteria use the Embden-Mayerhoff-Parnas pathway to produce fermentation products, whereas the Obscuribacterales representative is capable of using both aerobic and anaerobic respiration to produce energy.

(**Figure 2.2**). Under oxic conditions, we predict that *'O. phosphatis'* will fully oxidise one or more of these substrates via the TCA cycle, feeding NADH into the electron transport chain with a cbb3-type cytochrome as the terminal oxidase. This family of cytochromes is typically used in microaerophilic conditions [58] suggesting that *'O. phosphatis'* may be found within flocs where oxygen concentrations are lower [39]. Under anoxic conditions, we predict that it performs either respiration with nitrate as the terminal electron acceptor or, in the absence of nitrate, mixed-acid fermentation with the potential to produce ethanol, lactate, formate, succinate, CO_2 and H_2 (**Figure 2.2**). The presence of these metabolic pathways suggests that '*O. phosphatis*' has adapted to more dynamic environments (requiring greater metabolic plasticity) with 'feast-famine' nutrient cycles such as those artificially imposed on EBPR bioreactors.

Candidatus Caenarcanum bioreactoricola ('Caen.arc.an.um' L. neut. n. *caenum* mud, sludge; L. neut. n. *arcanum* secret, hidden; N.L. neut. n. *Caenarcanum* a bacterium hidden in sludge. 'bio.re.ac.to.ri.co.la'. L. suffix *-cola* inhabitant, dweller; N.L. masc. n. *bioreactericola* living in a bioreactor) UASB_169 (*'C. bioreactoricola'*) representing the order Caenarcaniphilales, has an estimated genome size of ~2 Mb and a remarkably low GC content of 27.7%, the lowest GC content reported for Cyanobacteria. Similar to the Gastranaerophilales genomes, '*C. bioreactoricola*' lacks the genes necessary for aerobic and anaerobic respiration, as well as the TCA cycle, suggesting that *'C. bioreactoricola'* has a streamlined metabolism only producing energy via fermentation with ethanol and lactate as the main fermentation products. *'C. bioreactoricola'* contains the subunits for Fe-only hydrogenase and the potential to produce hydrogen as a by-product from the fermentation process. Like the Gastranaerophilales, *'C. bioreactoricola'* may also be a hydrogen producer living in syntrophy with methanogens or acetogens in the bioreactor, as microcolonies of syntrophic bacteria are often observed in the granules from UASB systems and electron transfer in these microcolonies is thought to mostly occur through interspecies hydrogen transfer [59].

2.4.4 Emergence of photosynthesis in the Cyanobacteria

Melainabacteria resemble Oxyphotobacteria in their cell envelope gene complement comprising genes indicative of a Gram-negative (diderm) cell wall (**Figure 2.3**). This includes genes for the biosynthesis of Lipid A for the production of lipopolysaccharide (LPS) as previously reported for members of the Gastranaerophilales [12]. Oxyphotobacteria also have unusual cell envelope components for Gram-negative bacteria including porins (*somA, somB*), which are thought to help anchor the outer membrane to the peptidoglycan layer [60, 61]. All Melainabacteria have closely

related homologs to the oxyphotobacterial *somA* and *somB* genes suggesting that their cell envelopes comprise similar porins.

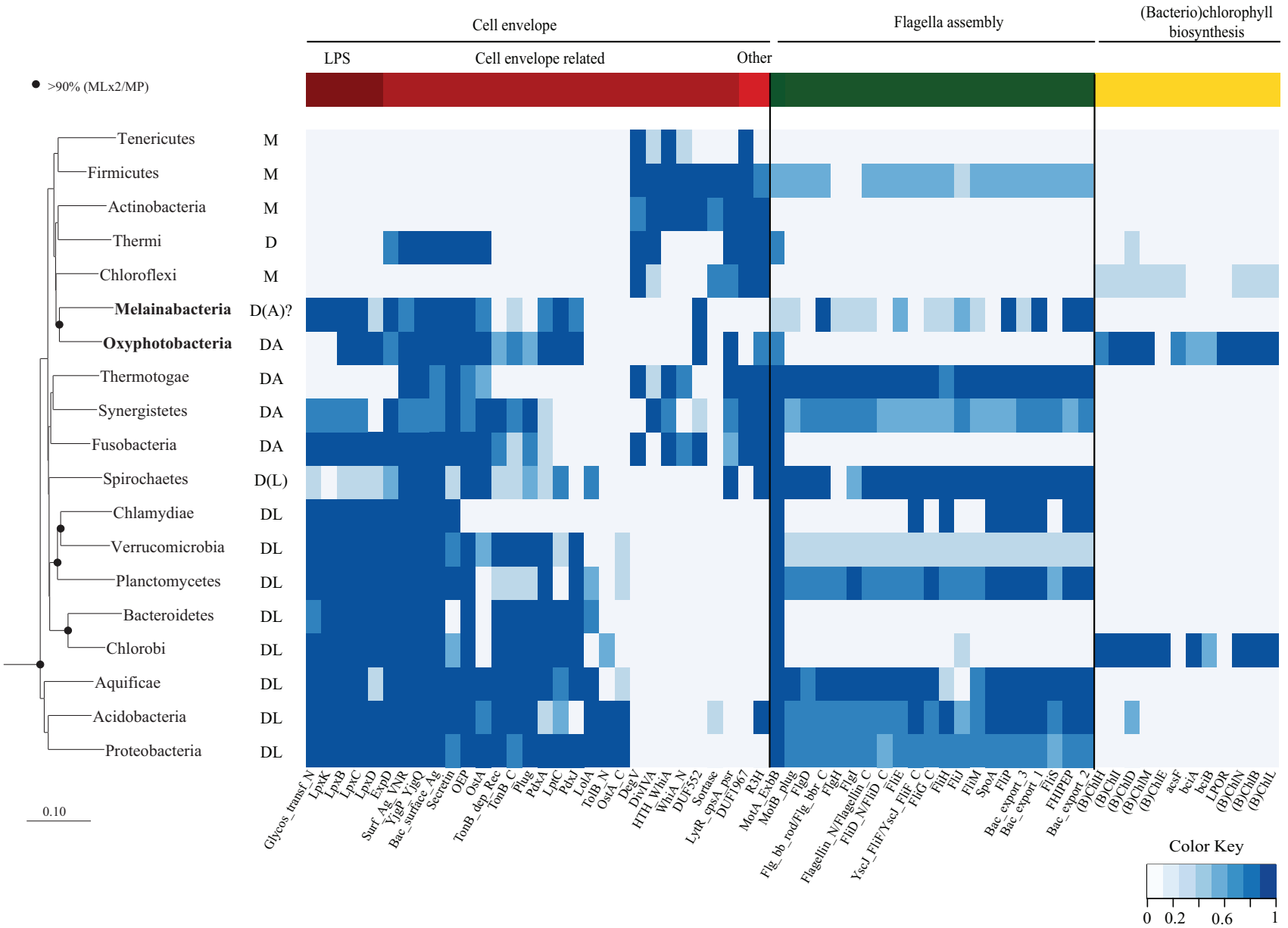
Di Rienzi *et. al*. highlighted the presence of putative circadian rhythm (*rpaA* and *rpaB*) and light response (*nblS*) regulators in the Gastranaerophilales, which are diagnostic of Cyanobacteria. We identified orthologs of these genes in all three orders of the Melainabacteria suggesting that these are uniquely ancestral features of the phylum. Together with the unambiguous phylogenetic placement of the Melainabacteria within the cyanobacterial radiation based on comparative analysis of highly conserved marker genes (**Figure 2.1**), the conservation of features characteristic of Oxyphotobacteria in the Melainabacteria further support a common ancestry and the proposal for a single phylum.

The most conspicuous difference between the Melainabacteria and Oxyphotobacteria is the absence of chlorophyll biosynthesis genes in the former (**Figure 2.3**) consistent with previous findings [12]. All subunits for photosystems I and II, and the electron transport chain were absent from the Melainabacteria genomes sequenced in this study. Another trait characteristic of photosynthetic cyanobacteria, carbon fixation, is similarly absent in the Melainabacteria (**Figure 2.2**) indicating that these organisms do not engage in a photoautotrophic lifestyle. Instead, members of this lineage appear to be chemoheterotrophs with diverse functionality.

The idea of non-photosynthetic Cyanobacteria is contrary to the prevailing dogma that all members of this phylum are photosynthetic [51]. However, this should not be a controversial conclusion given that photosynthesis is only found in sub-lineages of several other phyla such as the Proteobacteria, Firmicutes, Acidobacteria and Chloroflexi [62]. The complete absence of photosynthetic apparatus in the Melainabacteria suggests that the Oxyphotobacteria acquired photosystems after diverging from the common ancestor of the Melainabacteria (**Figure 2.1**). This is consistent with the inferences that photosynthesis genes have an extensive history of lateral transfer [2] and that photosynthesis developed late in the Cyanobacteria [63].

The acquisition of oxygenic photosynthesis in the Oxyphotobacteria had profound impact not only on the biosphere [2] but left imprints in their genomes that are now apparent by contrasting with their newly sequenced non-photosynthetic relatives. For example, there was a great expansion of ATP-driven transport systems in the Oxyphotobacteria likely for acquiring bicarbonate (COG0600/0715/1116) and iron (COG0609/1629/0735) necessary for photosynthesis and respiration (**Figure S2.9**). The additional energy available to Oxyphotobacteria via oxygenic

Figure 2.3. Distribution of key traits across the Cyanobacteria and other bacteria phyla



A maximum likelihood genome tree of the bacterial domain constructed using a concatenated alignment of 38 conserved proteins is shown at the left of the figure for phylogenetic ordering of traits shown in the heat map to the right. Black circles on interior nodes represent affiliations with >90% bootstrap support. Columns in the heat map represent individual gene families (Pfams or GIs; **Table S2.6**) grouped into three subsystems of interest; cell envelope, flagella and (bacterio)chlorophyll biosynthesis. Increasing representation of each gene family in a given phylum (percentage of genomes) is shown by increasing depth of color. Cell envelope classification is indicated by the abbreviations to the right of the phylum names: Monoderm (M), Diderm (D), Diderm-LPS (DL), Diderm-Atypical (DA) from Albertsen *et al.,* 2013. Note that it is not possible to determine if the Melainabacteria are diderms or atypical diderms based on sequence data only.

photosynthesis may also explain the widespread acquisition of energy-intensive biosynthetic pathways such as secondary metabolite synthesis [51] (**Figure S2.10**).

2.5 Conclusion

Our findings expand the recognised phylogenomic boundaries of the phylum Cyanobacteria. We infer that the cyanobacterial ancestor was a non-photosynthetic chemoheterotroph and that photosystems were acquired after divergence of the classes Melainabacteria and Oxyphotobacteria. We suggest that the acquisition of oxygenic photosynthesis resulted in an increase in genome complexity within the Oxyphotobacteria (followed by a subsequent reduction and streamlining in the *Prochlorococcus* lineage; [64]) while the Melainabacteria mostly retained a simpler ancestral metabolism. Consistent with the phylogenetic depth of the Melainabacteria, members of this class occupy a wide range of environmental niches with varied metabolic properties, mostly centred around fermentative lifestyles, that nonetheless extend the known metabolic diversity of the Cyanobacteria. These include respiratory nitrate reduction (*'O. phosphatis'*) and flagella-based motility (in some Gastranaerophilales [12]). If the inclusion of *Vampirovibrio chlorellavorus* [65] in the Melainabacteria is confirmed by genomic sequencing, then parasitism can also be added to the known phenotypes of Cyanobacteria. The availability of eleven high quality draft genomes representing multiple orders within the Melainabacteria (**Table 2.1**) provides a sound basis for further investigations into this fascinating group, for example, via spatial visualisation [66] and genome-directed isolation [67, 68].

2.6 Acknowledgements

We thank Norman Pace, Tal Dagan and Michael Galperin for providing valuable perspective on the study and Karen Nilsson and Jacqui Brumm at Lone Pine Koala Sanctuary for facilitating collection of koala faecal samples, Adam Skarshewski for implementing the concatenation of single copy marker genes, Serene Low and Margaret Butler for preparing samples for Illumina sequencing, Fiona May for 454 pyrotags, Queensland Centre for Medical Genomics, UQ and Aalborg University for Illumina paired end and mate pair sequencing. We also thank Akiko Ohashi and Satoko Matsukura at AIST for Illumina shotgun and 16S amplicon sequencing. We thank Tim Lilburn for information on the ATCC strain of *Vampirovibrio chlorellavorus*. We thank Satoshi Hanada and Aharon Oren for etymological advice and clarification of nomenclature versus classification issues (we are classifying here). This work is supported by the Australian Research Council (ARC) through project DP120103498, strategic funds from the Australian Centre for Ecogenomics; G.W.T. is supported by an ARC Queen Elizabeth II fellowship [DP1093175]; R.M.S and C.T.S is supported by an Australian Postgraduate Award (APA).

2.7 References

1. Nelson, N. and A. Ben-Shem, The complex architecture of oxygenic photosynthesis. Nature Reviews Molecular Cell Biology, 2004. **5**(12): p. 971-982.
2. Hohmann-Marriott, M.F. and R.E. Blankenship, Evolution of Photosynthesis. Annual Review of Plant Biology, 2011. **62**(1): p. 515-548.
3. Dojka, M.A., J.K. Harris, and N.R. Pace, Expanding the Known Diversity and Environmental Distribution of an Uncultured Phylogenetic Division of Bacteria. Applied and Environmental Microbiology, 2000. **66**(4): p. 1617-1621.
4. Rippka, R., et al., Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. Journal of General Microbiology, 1979. **111**(1): p. 1-61.
5. McDonald, D., et al., An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J, 2012. **6**(3): p. 610-8.
6. Quast, C., et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 2013. **41**(D1): p. D590-D596.
7. Williams, M.M., et al., Phylogenetic diversity of drinking water bacteria in a distribution system simulator. Journal of Applied Microbiology, 2004. **96**(5): p. 954-964.
8. Cruz-Martinez, K., et al., Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. ISME J, 2009. **3**(6): p. 738-44.
9. Liu, B., et al., *Thauera* and *Azoarcus* as functionally important genera in a denitrifying quinoline-removal bioreactor as revealed by microbial community structure comparison. FEMS Microbiology Ecology, 2006. **55**(2): p. 274-86.
10. Ley, R.E., et al., Microbial ecology: Human gut microbes associated with obesity. Nature, 2006. **444**(7122): p. 1022-1023.
11. Gromov, B. and K. Mamkaeva, Proposal of a new genus *Vampirovibrio* for chlorellavorus bacteria previously assigned to *Bdellovibrio*. Mikrobiologia, 1980. **49**: p. 165-167.
12. Di Rienzi, S.C., et al., The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. eLife, 2013. **2**: p. e01102.
13. Lu, H., et al., Obtaining highly enriched cultures of *Candidatus* Accumulibacter phosphates through alternating carbon sources. Water Research, 2006. **40**(20): p. 3838-48.
14. Matsuki, T., et al., Development of 16S rRNA-Gene-Targeted Group-Specific Primers for the Detection and Identification of Predominant Bacteria in Human Feces. Applied and Environmental Microbiology, 2002. **68**(11): p. 5445-5451.
15. Caporaso, J.G., et al., QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 2010. **7**(5): p. 335-336.

16. Bragg, L., et al., Fast, accurate error-correction of amplicon pyrosequences using Acacia. Nature Methods, 2012. **9**(5): p. 425-426.
17. Johnson, M., et al., NCBI BLAST: a better web interface. Nucleic Acids Research, 2008. **36**: p. W5-9.
18. DeSantis, T.Z., et al., Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology, 2006. **72**(7): p. 5069-72.
19. Caporaso, J.G., et al., Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J, 2012. **6**(8): p. 1621-1624.
20. Camacho, C., et al., BLAST+: architecture and applications. BMC Bioinformatics, 2009. **10**(1): p. 421.
21. Qin, J., et al., A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2010. **464**(7285): p. 59-65.
22. Imelfort, M., et al., GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ, 2014. **2**: p. e603.
23. Dupont, C.L., et al., Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J, 2012. **6**(6): p. 1186-99.
24. Finn, R.D., J. Clements, and S.R. Eddy, HMMER web server: interactive sequence similarity searching. Nucleic Acids Research, 2011. **39**: p. W29-37.
25. Punta, M., et al., The Pfam protein families database. Nucleic Acids Research, 2012. **40**(D1): p. D290-D301.
26. Haft, D.H., J.D. Selengut, and O. White, The TIGRFAMs database of protein families. Nucleic Acids Research, 2003. **31**(1): p. 371-3.
27. Parks, D.H., et al., CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. PeerJ PrePrints, 2014. **2**: p. e554v1.
28. Boetzer, M., et al., Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 2011. **27**(4): p. 578-9.
29. Li, H. and R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 2010. **26**(5): p. 589-95.
30. Ludwig, W., et al., ARB: a software environment for sequence data. Nucleic Acids Research, 2004. **32**(4): p. 1363-1371.
31. Swofford, D., PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4. 2003.
32. Darling, A.E., et al., PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ, 2014. **2**: p. e243.

33. Felsenstein, J., PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics, 1989. **5**: p. 164-166.
34. Markowitz, V.M., et al., IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Research, 2012. **40**(D1): p. D115-D122.
35. Price, M.N., P.S. Dehal, and A.P. Arkin, FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE, 2010. **5**(3): p. e9490.
36. Castresana, J., Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution, 2000. **17**(4): p. 540-52.
37. Stamatakis, A., RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 2006. **22**(21): p. 2688-90.
38. Markowitz, V.M., et al., IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics, 2009. **25**(17): p. 2271-2278.
39. Albertsen, M., et al., Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature Biotechnology, 2013. **31**(6): p. 533-538.
40. Pallen, M.J. and N.J. Matzke, From The Origin of Species to the origin of bacterial flagella. Nature Reviews Microbiology, 2006. **4**(10): p. 784-790.
41. Sousa, F.L., et al., Chlorophyll biosynthesis gene evolution indicates photosystem gene duplication, not photosystem merger, at the origin of oxygenic photosynthesis. Genome Biology and Evolution, 2013. **5**(1): p. 200-216.
42. Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 1997. **25**(17): p. 3389-402.
43. Racine, J.S., RStudio: A Platform-Independent IDE for R and Sweave. Journal of Applied Econometrics, 2012. **27**(1): p. 167-172.
44. Warnes, G.R., et al., gplots: Various R programming tools for plotting data. 2013.
45. Neuwirth, E., RColorBrewer: ColorBrewer palettes. 2011.
46. Hyatt, D., et al., Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 2010. **11**(1): p. 119.
47. Tatusov, R.L., et al., The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 2003. **4**: p. 41.
48. Parks, D.H., et al., STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics, 2014. **30**(21): p. 3123-4.
49. Sharon, I., et al., Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Research, 2013. **23**(1): p. 111-20.

50. Hugenholtz, P., B.M. Goebel, and N.R. Pace, Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology, 1998. **180**(18): p. 4765-74.
51. Shih, P.M., et al., Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(3): p. 1053-8.
52. Wagner, M. and M. Horn, The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Current Opinion in Biotechnology, 2006. **17**(3): p. 241-9.
53. Gibbons, N.E. and R.G.E. Murray, Validation of *Cyanobacteriales* Stanier in Gibbons and Murray 1978 as a New Order of the Kingdom *Procaryotae* Murray 1968, and of the Use of Neuter Plural Endings for *Photobacteria* and *Scotobacteria* classes nov. Gibbons and Murray 1978: Request for an Opinion. International Journal of Systematic Bacteriology, 1978. **28**(2): p. 332-333.
54. Komárek, J., Recent changes (2008) in cyanobacteria taxonomy based on a combination of molecular background with phenotype and ecological consequences (genus and species concept). Hydrobiologia, 2010. **639**(1): p. 245-259.
55. Quigley, E.M. and R. Quera, Small intestinal bacterial overgrowth: roles of antibiotics, prebiotics, and probiotics. Gastroenterology, 2006. **130**: p. S78-90.
56. Goris, J., et al., DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology, 2007. **57**: p. 81-91.
57. Blackall, L.L., et al., A review and update of the microbiology of enhanced biological phosphorus removal in wastewater treatment plants. Antonie Van Leeuwenhoek, 2002. **81**(1-4): p. 681-91.
58. Kulajta, C., et al., Multi-step assembly pathway of the *cbb*$_3$-type cytochrome c oxidase complex. Journal of Molecular Biology, 2006. **355**(5): p. 989-1004.
59. Schmidt, J.E. and B.K. Ahring, Granular sludge formation in upflow anaerobic sludge blanket (UASB) reactors. Biotechnology and Bioengineering, 1996. **49**(3): p. 229-46.
60. Hansel, A., et al., Cloning and characterization of the genes coding for two porins in the unicellular cyanobacterium *Synechococcus* PCC 6301. Biochimica et Biophysica Acta, 1998. **1399**(1): p. 31-9.
61. Hoiczyk, E. and A. Hansel, Cyanobacterial Cell Walls: News from an Unusual Prokaryotic Envelope. Journal of Bacteriology, 2000. **182**(5): p. 1191-1199.

62. Bryant, D.A. and N.U. Frigaard, Prokaryotic photosynthesis and phototrophy illuminated. Trends in Microbiology, 2006. **14**(11): p. 488-96.
63. Xiong, J., et al., Molecular Evidence for the Early Evolution of Photosynthesis. Science, 2000. **289**(5485): p. 1724-1730.
64. Partensky, F. and L. Garczarek, *Prochlorococcus*: advantages and limits of minimalism. Annual Review of Marine Science, 2010. **2**: p. 305-31.
65. Coder, D.M. and L.J. Goff, The host range of the Chlorellavorus bacterium ("*Vampirovibrio chlorellavorus*"). Journal of Phycology, 1986. **22**(4): p. 543-546.
66. Moter, A. and U.B. Göbel, Fluorescence in situ hybridization (FISH) for direct visualization of microorganisms. Journal of Microbiological Methods, 2000. **41**(2): p. 85-112.
67. Pope, P.B., et al., Isolation of *Succinivibrionaceae* implicated in low methane emissions from Tammar wallabies. Science, 2011. **333**(6042): p. 646-8.
68. Tyson, G.W., et al., Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community. Applied and Environmental Microbiology, 2005. **71**(10): p. 6319-24.

Chapter 3: Back from the dead; the curious tale of the predatory cyanobacterium *Vampirovibrio chlorellavorus*

Rochelle M. Soo[1], Ben J. Woodcroft[1], Donovan H. Parks[1], Gene W. Tyson[1,2] and Philip Hugenholtz[1,3]*

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia

[2]Advanced Water Management Centre, The University of Queensland, St Lucia, QLD 4072, Australia

[3]Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia

3.1 Abstract

An uncultured non-photosynthetic basal lineage of the Cyanobacteria, the Melainabacteria, was recently characterised by metagenomic analyses of aphotic environmental samples. However, a predatory bacterium, *Vampirovibrio chlorellavorus,* originally described in 1972 appears to be the first cultured representative of the Melainabacteria based on a 16S rRNA sequence recovered from a lyophilised co-culture of the organism. Here, we sequenced the genome of *V. chlorellavorus* directly from 36 year-old lyophilised material that could not be resuscitated confirming its identity as a member of the Melainabacteria. We identified attributes in the genome that likely allow *V. chlorellavorus* to function as an obligate predator of the microalga *Chlorella vulgaris,* and predict that it is the first described predator to use an *Agrobacterium tumefaciens*-like conjugative type IV secretion system to invade its host. *V. chlorellavorus* is the first cyanobacterium recognised to have a predatory lifestyle and further supports the assertion that Melainabacteria are non-photosynthetic.

3.2 Introduction

Predatory microorganisms attack and digest their prey, which can be either bacteria or microbial eukaryotes [1, 2]. They have been found in a range of environments, including terrestrial, freshwater, estuaries, oceans, sewages and animal faeces [3]. Microbial predators have been classified as obligate (unable to grow in the absence of prey) or facultative (able to grow as a pure culture without the presence of prey). In addition they can be periplasmic (penetrate and attach to the inner membrane), epibiotic (attach to the outside), endobiotic (penetrate the cytoplasm) or wolf-pack (swarming as a

'wolf-pack' towards prey, which they kill and degrade) [4, 5]. To date, four bacterial phyla harbour microbial predators; the Proteobacteria, Actinobacteria, Bacteroidetes and Chloroflexi [2, 6, 7].

In 1972, Gromov and Mamkaeva first described the predatory nature of *Bdellovibrio chlorellavorus* towards the microalgae *Chlorella vulgaris* in a Ukrainian freshwater reservoir [8]. They reported that co-inoculation of the alga and bacterium resulted in clumping and colour change of algal cells, formation of refractile bodies and finally algal cell death. However, unlike other *Bdellovibrio* species that invade the periplasm of Gram-negative bacteria, *B. chlorellavorus* only attached to the surface of *C. vulgaris*, producing peripheral vacuoles in the alga followed by a gradual dissolution of the infected cell contents [9]. This distinct mode of predation called into question the classification of *B. chlorellavorus* as a *Bdellovibrio* [1] resulting in its reclassification as *Vampirovibrio chlorellavorus* in 1980, although its higher level assignment to the Deltaproteobacteria was retained [10].

Co-cultures of *V. chlorellavorus* and *C. vulgaris* were deposited in three culture collections in 1978 [1]. However, to the best of our knowledge there are no reports of successful resuscitation of the organism from lyophilised material. The only subsequent studies of *V. chlorellavorus* were based on co-cultures obtained directly from the investigators who originally enriched the bacterium [9, 11]. The American Type Culture Collection (ATCC) was able to successfully extract DNA from one of the 32 year-old lyophilised co-cultures and sequence the 16S rRNA gene of *V. chlorellavorus* (Genbank acc. no. HM038000). Comparative analyses of this sequence indicate that *V. chlorellavorus* is actually a member of the phylum Cyanobacteria rather than the Proteobacteria according to the Greengenes [12] and Silva [13] taxonomies. This may explain why the culture could not be revived as Cyanobacteria are notoriously difficult to resuscitate from lyophilised material [14]. More specifically, *V. chlorellavorus* is a member of a recently described basal lineage of non-photosynthetic Cyanobacteria, the class Melainabacteria ([15], originally classified as a separate phylum [16]). Here, we report the near-complete genome of *V. chlorellavorus* sequenced directly from a 36-year-old vial of co-cultured lyophilised cells, confirm its phylogenetic position in the Cyanobacteria, and infer the molecular underpinnings of its predatory life cycle.

3.3 Materials and Methods

3.3.1 Sample collection

Co-cultured *Vampirovibrio chlorellavorus* and *Chlorella vulgaris* (NCIB 11383) (deposited in 1978 by Coder and Starr) were obtained as lyophilised cells from the National Collections of Industrial, Food and Marine Bacteria (NCIMB), Aberdeen, Scotland.

3.3.2 Genomic DNA extraction

Genomic DNA (gDNA) was extracted from lyophilised cells using a MoBio Soil Extraction kit (MoBio Laboratories, Carlsbad, CA). gDNA was quantified using a Qubit 2.0 fluorometer (Life technologies). 1ng of the gDNA was used to construct a paired-end library with the Illumina Nextera XT DNA Sample Preparation kit according to protocol but with double size selection to obtain an insert size of 300-800 bp [17]. The library was sequenced on an Illumina Miseq system using the Miseq Reagent Kit v3 at the Institute of Molecular Bioscience, University of Queensland.

3.3.3 Genome assembly, completeness and contamination

Sequencing reads were processed with FastQC to check for quality (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Illumina Nextera adaptors were removed using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Reads were parsed through GraftM (https://github.com/geronimp/graftM) version r2439db using the May, 2013 version of the Greengenes database 97% OTUs (operational taxonomic units) as a reference [12] to identify those containing parts of 16S or 18S rRNA genes using default parameters. The 5' end of all reads was trimmed (~20bp) to remove low-quality sequence and paired reads were assembled into contigs with a kmer size of 63 using CLC Genomics Workbench v7.0 (CLC bio). The statistical package R with ggplot2 (https://github.com/hadley/ggplot2) was used to plot GC content against coverage allowing contigs belonging to the *V. chlorellavorus* genome to be identified. A discrete cluster of contigs with >180x coverage and a GC range of 42-54% were identified as belonging to *V. chlorellavorus*, while contigs with <180x coverage were assigned to *C. vulgaris* (**Figure S3.1**). BLASTN [18] (v2.2.29+) using default settings was used to verify that contigs with >180x coverage had homology to bacterial sequences with NCBI's non-redundant database. Additionally, the 16S rRNA gene was identified using Prokka v1.8 [19] and a BLASTN search was used to identify the closest neighbour in the May, 2013 version of the Greengenes database [12]. The completeness and contamination of the genome belonging to *V. chlorellavorus* was examined using CheckM v0.9.5 [20] with a set of 104 conserved bacterial single-copy marker genes [15]. IslandViewer was used to identify genomic islands [21] with the SIGI-HMM programme [22].

Plasmids were identified using the 'roundup' mode of FinishM git version 5664703 (https://github.com/wwood/finishm), using raw reads as input, a kmer length of 51bp and a coverage cutoff of 15. A combination of manual inspection of the assembly graph generated using the 'visualise' mode and automated assembly with the 'assemble' mode confirmed that the contig ends unambiguously joined together (i.e. they joined together and to no other contig ends) and that the two plasmid contigs originally assembled with CLC were otherwise free of mis-assemblies. Plasmids

were also confirmed by the annotation of multiple transfer (*tra*) genes by the Integrated Microbial Genomics Expert Review (IMG/ER) system (see below).

3.3.4 Genome annotation

The *V. chlorellavorus* genome was submitted to IMG/ER for annotation [23]. The *V. chlorellavorus* genome has been deposited at JGI [JGI IMG-ER:2600254900]. The genome was also annotated with Prokka v.1.8 [19] and the Uniref 90 database [24]. KEGG maps [25] and gene annotations were used to reconstruct the metabolism of the *V. chlorellavorus* genome. Individual genes that were annotated as 'hypothetical protein' or had been potentially misannotated based on the annotation of surrounding genes were further explored through BLASTN searches against the NCBI-nr database. A metabolic cartoon was prepared in Adobe Illustrator CS6.

The methyl-accepting chemotaxis proteins identified by IMG-ER were submitted to InterProScan5 [26] to determine chemotaxis protein domains. Putative genes were annotated with the dbCAN web server [27] to identify glycoside hydrolases and checked against the IMG annotations and BLAST results. The MEROPS server [28] was used to identify putative peptidases in *V. chlorellavorus* using batch BLAST.

A Genbank file for *V. chlorellavorus* was generated through the xBASE website [29]. The ribosomal proteins, chaperones and transcriptional and translational proteins of *V. chlorellavorus* were used as representatives of recognised highly expressed genes to identify other putatively highly expressed genes in the genome using PHX (predicted highly expressed) analysis using the standard genetic code (http://www.cmbl.uga.edu/software/phxpa.html; [30, 31]. Putatively horizontally transferred (alien) genes were identified by their atypical codon usage from the genome average also using PHX analysis.

3.3.5 Phylogenetic tree

A bacterial genome tree was inferred in order to establish the phylogenetic relationship of the *V. chlorellavorus* genome. A set of 5,449 bacterial genomes previously identified as being of exceptional quality were used to establish a set of bacterial marker genes suitable for phylogenetic inference [20]. An initial set of 178 single copy genes present exactly once in >90% of the trusted genomes (found in >90% of the genomes) was identified using the PFAM [32] and TIGRFAM [33] annotations provided by the Integrated Microbial Genomes (IMG; [34]). The same protein family may be represented in both PFAM and TIGRFAM. Families from these two databases were considered redundant if they matched the same genes in >90% of the trusted genomes, in which case

preference was given to the TIGRFAM families. Genes present multiple times within a genome were considered to have congruent phylogenetic histories if all copies of the gene were situated within a single conspecific clade within its gene tree. From the 178 initial genes, 69 were removed from consideration as they exhibited divergent phylogenetic histories in >1% of the trusted genomes (**Table S3.1**). The remaining 109 genes were identified across an expanded set of 7732 bacterial genomes, including all known Melainabacteria genomes along with an outgroup of 169 archaeal genomes using Prodigal v2.60 [35] to identify call genes and HMMER v3.1b1 (http://hmmer.janelia.org) to assign genes to PFAM and TIGRFAM families. Gene assignment was performed using model specific cutoff values for both the PFAM (-cut_gc) and TIGRFAM (-cut_tc) HMMs. For both the individual gene trees and concatenated genome tree, genes were aligned with HMMER v3.1b1 and phylogenetic inference performed with FastTree v2.1.7 [36] under the WAG+GAMMA model. Support values for the bacterial genome tree were determined by applying FastTree to 100 bootstrapped replicates [37]. The 16S rRNA gene tree was constructed as previously described [15]. Briefly, the 16S rRNA gene from *V. chlorellavorus* was aligned to the standard Greengenes alignment with PyNAST [12]. Aligned sequences and a Greengenes reference alignment, version gg_13_5 were imported into ARB and the *V. chlorellavorus* sequence alignment was corrected using the ARB EDIT tool. Representative taxa (>1,300 nt) were selected for constructing the alignments, which were exported from ARB [38] with Lane mask filtering. Neighbour joining trees were calculated from the mask alingments with LogDet distance estimation using PAUP*4.0 [39] with 100 bootstrap replicates. Maximum parsimony trees were calculated using PAUP*4.0 [39] with 100 bootstrap replicates. Maximum likelihood trees were calculated from the masked alignments using the Generalised Time-Reversible model with Gamma and I options in RAxML version 7.7.8 [40] (raxmlHPC-PTHREADS -f a -k -x 12345 -p 12345 -N 100 -T 4 -m GTRGAMMAI). Bootstrap resampling data (100 replicates) were generated with SEQBOOT in the phylip package [41] and used for 100 bootstrap resamplings. Generated trees were re-imported into ARB for visualisation.

3.3.6 Phylogenetic trees for *virB4* and *fliI* genes

VirB4 sequences were obtained from Guglielmini et al. 2012. The phylip file (figure3_mafft_alignment.phy) obtained from the DRYAD database was converted to an HMM using HMMer v3.1b1 (http://hmmer.janelia.org) and the VirB4 sequences from *V. chlorellavorus* was aligned to the HMM. The aligned sequences were used to construct a phylogenetic tree with phyml (v3.1) [42] using default settings [43].

The HMM for TIGR03496 (FliI_clade 1) was used to identify *fliI* genes from 2,256 finished genomes in the IMG database v4 and the 12 Melainabacteria genomes, including *V. chlorellavorus*. A phylogenetic tree of the *fliI* genes was constructed using FastTree (version 2.1.7) with default settings [36].

3.3.7 Comparison of *V. chlorellavorus* to other predatory bacteria

The presence of orthologues for differentiating predatory and non-predatory bacteria as described in Pasternak et al. (2013) were identified in the *V. chlorellavorus* genome using BLASTP [18] against the OrthoMCL DB v4 [44] with an e-value threshold of 1e-5.

3.3.8 Comparison of *V. chlorellavorus* to other Melainabacteria genomes

Eleven Melainabacteria genomes were compared to the *V. chlorellavorus* genome [15, 16]. COG profiles were constructed using homology search between putative genes predicted with Prodigal v2.60 [45] and the 2003 COG database [46]. Genes were assigned to COGs using BLASTP (v2.2.22) with an e-value threshold of 1e-2, an alignment length threshold of 70% and a percent identity threshold of 30%. The relative percentage of a COG category was calculated in relation to the total number of putative genes predicted for each genome. STAMP v2.0.8 [47] was used to explore the resulting COG profiles and create summary plots.

3.4 Results and Discussion

3.4.1 Genome summary

A total of 701.2 Mbp of shotgun sequence data (2 x 300 bp paired-end Illumina) was obtained from DNA extracted from a co-culture of *Vampirovibrio chlorellavorus* and *Chlorella vulgaris* (NCIB 11384). A search of the unassembled dataset for 16S rRNA sequences revealed 333 reads mapping to *V. chlorellavorus* (16 chloroplast, 3 mitochondria). No matches to other microorganisms were identified. Sequence reads were assembled into 113 contigs comprising 3.2 Mbp. Ordination of the data by GC content and mapping read depth revealed a high coverage cluster of contigs comprising ~94% of the data (**Figure S3.1**). These contigs were inferred to belong to *V. chlorellavorus* by the presence of a 16S rRNA gene on one of the contigs (*see below*) and low coverage contigs were inferred to belong to the *C. vulgaris* by best matches to reference *Chlorella* genomes. Inspection of the assemblies showed no evidence for microheterogeneity (SNPs, indels) in the *V. chlorellavorus* contigs suggesting that it was a pure bacterial strain. After manual curation, the genome of *V. chlorellavorus* was represented by 26 contigs comprising a total of 2.91 Mbp with an average GC content of 51.4% and two plasmids comprising ~72 Kbp and ~50 Kbp were identified which

contained genes for conjugative gene transfer (*see below*). These plasmids had mapping coverage similar to the genomic contigs suggesting that they are low-copy. The genome was estimated to be near-complete with low contamination according to CheckM [20] suggesting that the fraction of missed genes in contig gaps was minimal. The protein coding density of the genome is 87.1% and predicted to encode 2,847 putative genes, 41 tRNA genes which represent all 20 amino acids and one rRNA operon (only the 16S and 23S rRNA genes were identified). Approximately two thirds (69.9%) of the putative genes can be assigned to a putative function and half (53.2%) can be assigned to a COG category. *V. chlorellavorus* contains 13 transposases and 18 genomic islands (genomic regions that are thought to have horizontal origins) (**Table 3.1**).

Table 3.1. Features of the *Vampirovibrio chlorellavorus* genome

Isolate name	*Vampirovibrio chlorellavorus*
Closest 16S rRNA gene clone[a]	HG-B02128 (JN409206)
Number of contigs	26
Number of plasmids	2
Total length (bp)	3,030,230
N50	217,646
GC (%)	51.4
tRNA genes	41
rRNA genes found in genome	16S, 23S
Putative genes	2,844
Genomic islands[b]	18
Mobile genetic elements	13 transposases
CDS coding for hydrolytic enzymes	106 proteases/peptidases
	0 DNases
	0 RNases
	0 glycanases
	3 lipases/esterases
	2 lysophospholipase
Genome completeness[c]	100% (104/104)
Genome contamination[c]	0.95% (1/104)
Proposed class	Melainabacteria
Proposed order	Vampirovibrionales

[a] BLASTN search was used to identify the closest neighbour in the May, 2013 version of the Greengenes database [12].

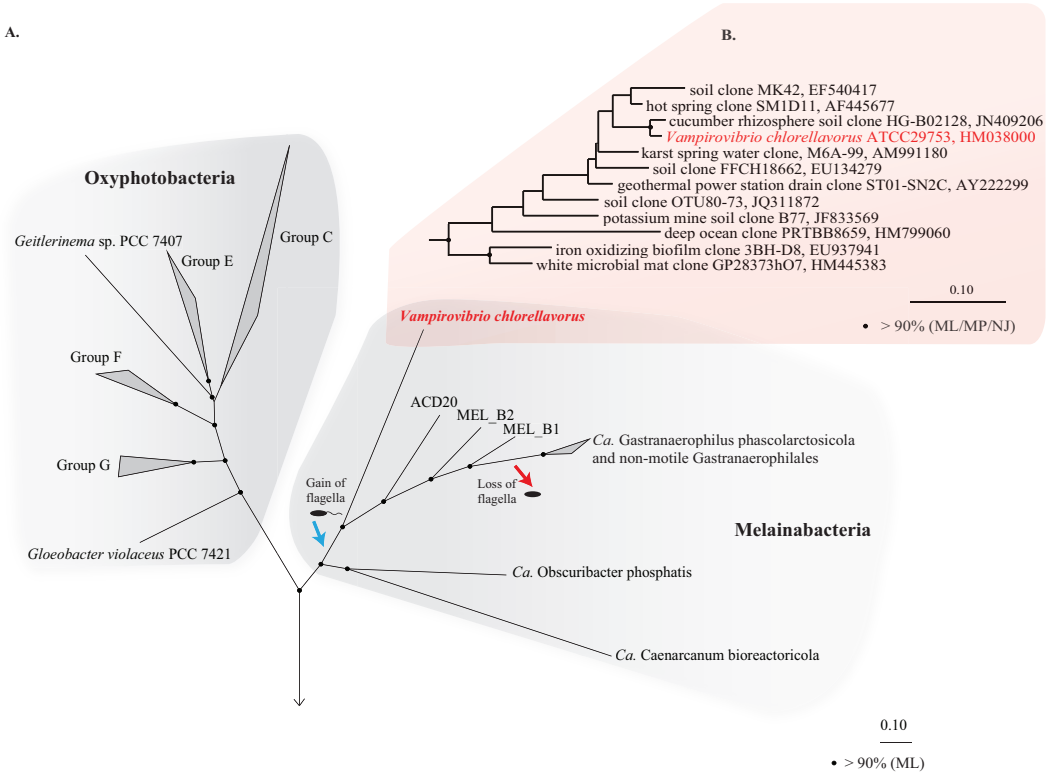
[b] IslandViewer was used to identify genomic islands [21] with the SIGI-HMM programme [22].

[c] The completeness and contamination of the population genome bin belonging to *V. chlorellavorus* was examined using CheckM v0.9.5 [20].

3.4.2 Phylogeny and taxonomy

The 16S rRNA gene obtained from the draft genome is identical to the reference sequence for *V. chlorellavorus* ATCC 29753 (acc. HM038000) and comparative analysis confirmed its placement as a deep-branching member of the Cyanobacteria phylum within the class Melainabacteria and order Vampirovibrionales ([15]; **Figure 3.1B**). Importantly, a concatenated gene tree of 109 conserved single copy genes produced a robust topology consistent with the 16S rRNA tree, also placing *V. chlorellavorus* in the class Melainabacteria (**Figure 3.1A**: **Figure S3.2**). These phylogenetic inferences clearly indicate that *V. chlorellavorus* is not a member of the Deltaproteobacteria as first suggested [8].

Figure 3.1. Phylogenetic position of *Vampirovibrio chlorellavorus* in the phylum Cyanobacteria



A) A maximum likelihood (ML) phylogenetic tree of the phylum Cyanobacteria inferred from a concatenated alignment of 109 single copy marker genes conserved across the bacterial domain. Black circles represent branch nodes with >90% bootstrap support by ML analysis. Class Oxyphotobacteria group names are according to Shih et al. 2013. The blue and red arrow indicate putative acquisition and loss of flagella respectively in the class Melainabacteria. Representatives of 32 bacterial phyla were used as outgroups in the analysis (**Figure S3.2**). *Ca* = *Candidatus*. **B**) A ML tree of the order Vampirovibrionales [15] based on aligned 16S rRNA gene sequences from the May, 2013 Greengenes database [12]. Black circles represent nodes with >90% ML, maximum parsimony (MP) and neighbour joining (NJ) bootstrap support values.

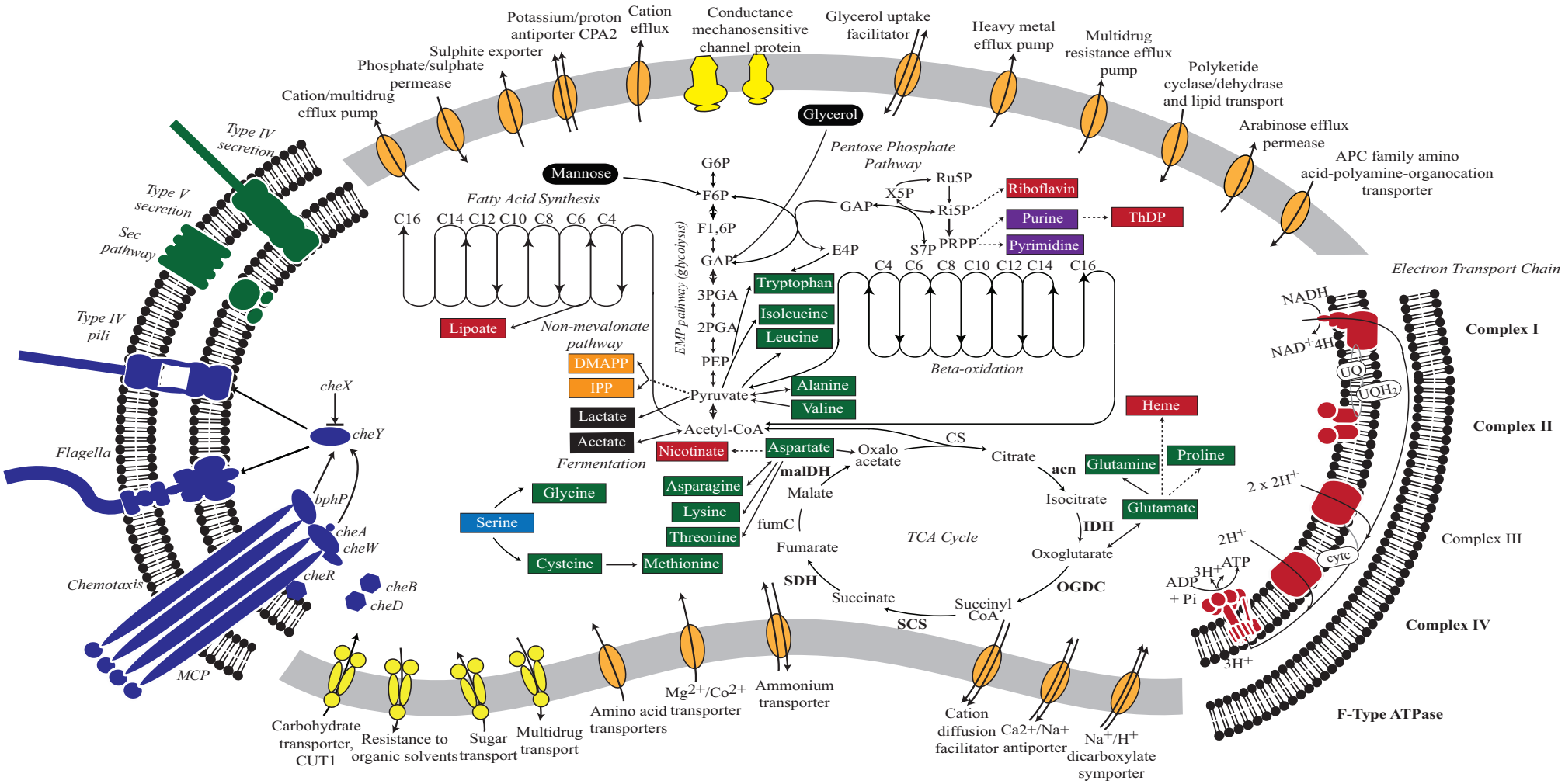
3.4.3 Cell shape and envelope

Microscopy studies revealed that *V. chlorellavorus* has a pleomorphic life cycle, being cocci during its free-living phase and vibrioid once attached to its host [1]. The *V. chlorellavorus* genome contains a gene that may function to spatially organise peptidoglycan synthesis (*mreB*) and a key cell division protein (*ftsZ*), which have been shown to be necessary for the maintenance of cell shape in *Caulobacter crescentus* and *Eschericia coli* [49, 50]. The bacterium also contains the genes indicative of a Gram-negative cell envelope including those for the production of lipopolysaccharide (LPS), Lipid A and O-antigen [51]. This is consistent with prior ultrastructural imaging of *V. chlorellavorus* which showed this bacterium has a typical Gram-negative cell envelope [1]. Interestingly, the genome also contains surface layer homology (SLH) domains, suggesting that the cell has the capacity to produce an S-layer, although no such structures were observable in transmission electron microscopy (TEM) images [1, 11]. This does not preclude their presence, however, because the samples were not processed optimally for S-layer visualisation; and under unfavourable laboratory cultivation conditions, the formation of the S-layer may be lost [52, 53]. S-layers have been observed in at least 60 strains of Cyanobacteria [53] and SLH domains have also been found in other Melainabacterial genomes.

3.4.4 Core metabolism

The *V. chlorellavorus* genome encodes a complete glycolysis pathway utilising glucose-6-phosphate, glycerol and mannose, the pentose phosphate pathway and a tricarboxylic acid (TCA) cycle. The genome also contains a complete set of genes for an electron transport chain comprising Complexes I to IV and an F-type ATPase. It has two terminal oxidases; a bd-type quinol and a cbb3-type cytochrome (Complex IV), both of which are used for microaerobic respiration [54]. According to PHX (predicted highly expressed) analysis [31], many of the genes in the glycolysis pathway, TCA cycle and electron transport chain are predicted to be highly expressed (**Figure 3.2; Table S3.2**) suggesting oxidative metabolism is central to the predatory lifestyle of *V. chlorellavorus* despite the inference of adaptation to low oxygen conditions. However, the genome also contains lactate dehydrogenase suggesting that it is able to ferment pyruvate to lactate under anaerobic conditions (**Figure 3.2**). The bacterium contains genes for fatty acid biosynthesis and β-oxidation, which leads to the production of acetyl-CoA. Consistent with other described members of the class Melainabacteria, and in contrast to oxygenic photosynthetic cyanobacteria, *V. chlorellavorus* lacks genes for photosynthesis and carbon fixation [15]. *V. chlorellavorus* can synthesise its own nucleotides and several cofactors and vitamins including lipoate, nicotinate, heme, riboflavin and

Figure 3.2. Metabolic reconstruction of *Vampirovibrio chlorellavorus*



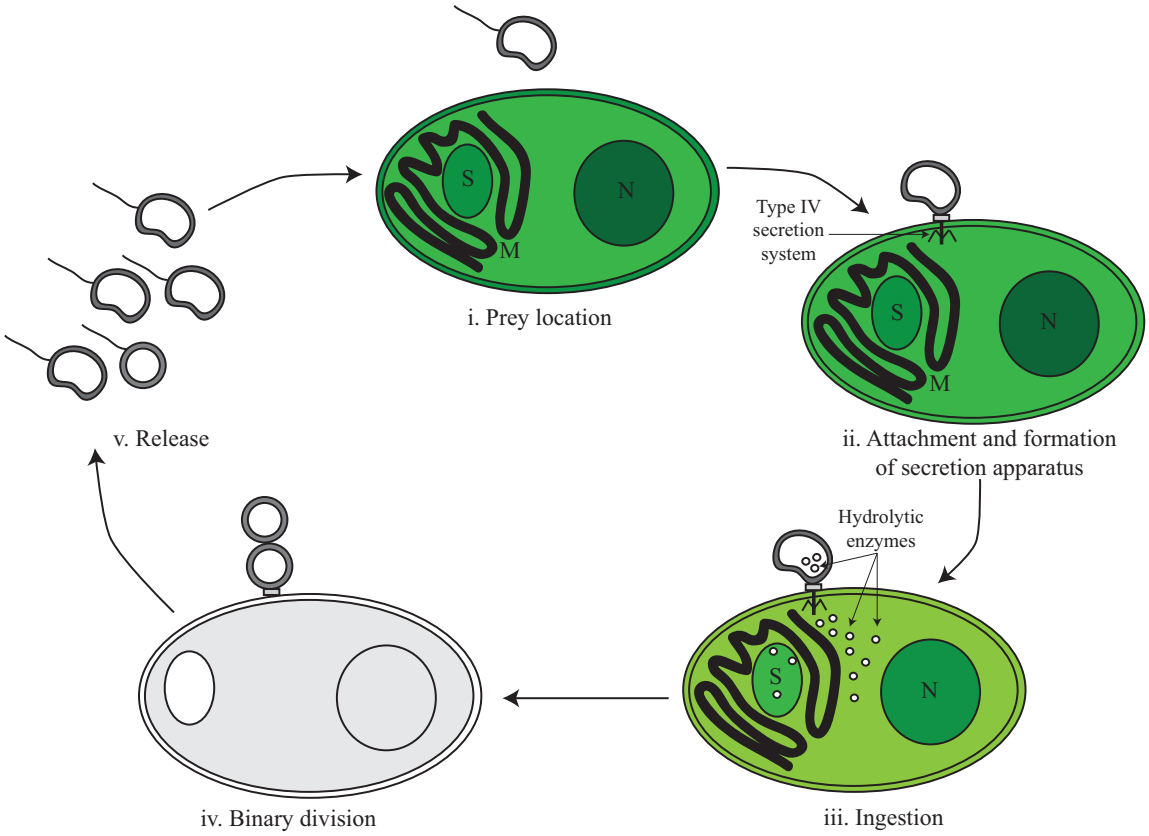
Metabolic predictions for *V. chlorellavorus* based on genes annotated by IMG/ER [23]. Solid and dashed lines represent single or multiple steps in a pathway respectively. Black ovals indicate substrates that enter the glycolysis pathway. Fermentation end-products are indicated as black rectangles. *V. chlorellavorus* is capable of oxidative phosphorylation as it contains a complete TCA cycle and electron transport chain. Biosynthetic products are shown in green (amino acids), red (co-factors and vitamins), purple (nucleotides), and orange (non-mevalonate pathway products). Serine (highlighted in blue) is not able to be synthesised and is presumably transported into the cell. ATP-binding cassette transporters are highlighted in yellow and permeases, pumps and transporters are highlighted in orange. The direction of substrate transport across the membrane is shown with arrows. Putatively highly expressed genes and complexes are bolded. *V. chlorellavorus* is missing all recognised photosynthesis genes including those for Photosystems I and II, chlorophyll and antennae proteins.

thiamine-diphosphate, but only 15 amino acids: alanine, asparagine, aspartate, cysteine, glutamate, glutamine, glycine, isoleucine, leucine, lysine, methionine, proline, threonine, tryptophan and valine. Although *V. chlorellavorus* does not have the genes necessary to synthesise the remaining five amino acids or their polyamine derivatives, it contains amino acid and polyamine transporters (**Figure 3.2**) that would allow it to obtain these organic compounds from external sources, most likely *C. vulgaris*.

3.4.5 The predatory lifestyle of *Vampirovibrio chlorellavorus*

Based on genomic inference and electron microscopy images obtained by Coder & Starr (1978), we divide the predatory life cycle of *V. chlorellavorus* into five phases comprising i) prey location, ii) attachment, iii) formation of secretion apparatus and ingestion, iv) binary division and v) release (**Figure 3.3**).

Figure 3.3. Proposed predatory life cycle of *Vampirovibrio chlorellavorus* informed by genome annotations



i) *V. chlorellavorus* seeks out *C. vulgaris* cells via chemotaxis and flagella. ii) it attaches to prey cells via a type IV secretion system (T4SS). iii) plasmid DNA and hydrolytic enzymes are transferred to the prey cells via the T4SS where they degrade algal cell contents (see **Figure 3.4** for details). iv) algal cell exudates are ingested by *V. chlorellavorus* allowing it to replicate by binary division. v) progeny are released completing the cycle. S - starch granule, M - mitochondria and N - nucleus.

3.4.5.1 Phase i: Prey location

The *V. chlorellavorus* genome encodes two-component regulatory systems including the well-known CheA-CheY signal transduction pathway that couples to flagella rotation or pili extension, attachment and retention (**Figure 3.2**) allowing the cell to move towards chemoattractants or away from chemorepellents [55]. Coder and Starr (1978) showed that *V. chlorellavorus* is able to swim towards its prey using a single, polar unsheathed flagellum possibly assisted by pili visible as thick bundles in proximity to the flagellum. All of the genes necessary to produce a functional flagellum and type IV pili (TFP) are present in the *V. chlorellavorus* genome ([56]; **Table S3.3**). In Cyanobacteria, *Synechocystis* strain PCC 6803 uses TFP for motility and it has also been speculated that TFP can drive motility in *Nostoc punctiforme* [30, 57]. It is likely that *V. chlorellavorus* uses chemotaxis to help it locate prey, but based on genome inference alone, it is not possible to determine which gradients *V. chlorellavorus* is detecting and responding to. However, the genome does contain one globin-coupled sensor inferred to be used for aerotaxis ([58]; **Figure S3.3**) and one putative light-activated kinase (bacteriophytochrome; [59]; BphP in **Figure 3.2**) that may enable *V. chlorellavorus* to move towards oxic and illuminated regions of its habitat that have a higher likelihood of containing *Chlorella* cells.

3.4.5.2 Phase ii: Attachment and formation of a conjugative secretion apparatus

V. chlorellavorus has a number of cellular features that likely facilitate its observed attachment to *Chlorella* cells: TFP (*described above*), an outer membrane protein (OmpA) and von Willebrand domain-containing proteins. While there are no reports of bacteria adhering to unicellular microbial eukaryotes using these structures, there are a number of examples for adherence to animal tissues. TFP are known to be involved in adhesion of pathogenic *Escherichia coli* and *Neisseria meningitidis* to human epithelial cells as a key virulence mechanism [60, 61]. OmpA porins are outer membrane proteins that assemble into an eight stranded β-barrel structure with four surface-exposed loops. Shin et al. (2005), showed that OmpA surface loops are critical for adhesion of *E. coli* to brain microvascular endothelial cells leading to neonatal meningitis [62]. Furthermore, OmpA is involved in the binding of *Acinetobacter baumanii* and *Pasteurella multocida* to fibronectin from human lung carcinoma [63]. The von Willebrand factor A (VWA) domains are found predominantly in cell adhesion and extracellular matrix molecules, including integrins, hemicentins and matrilins [64]. *Enterococcus faecalis* VWA domains are able to mediate protein-protein adhesion through a metal ion-dependent adhesion site [65].

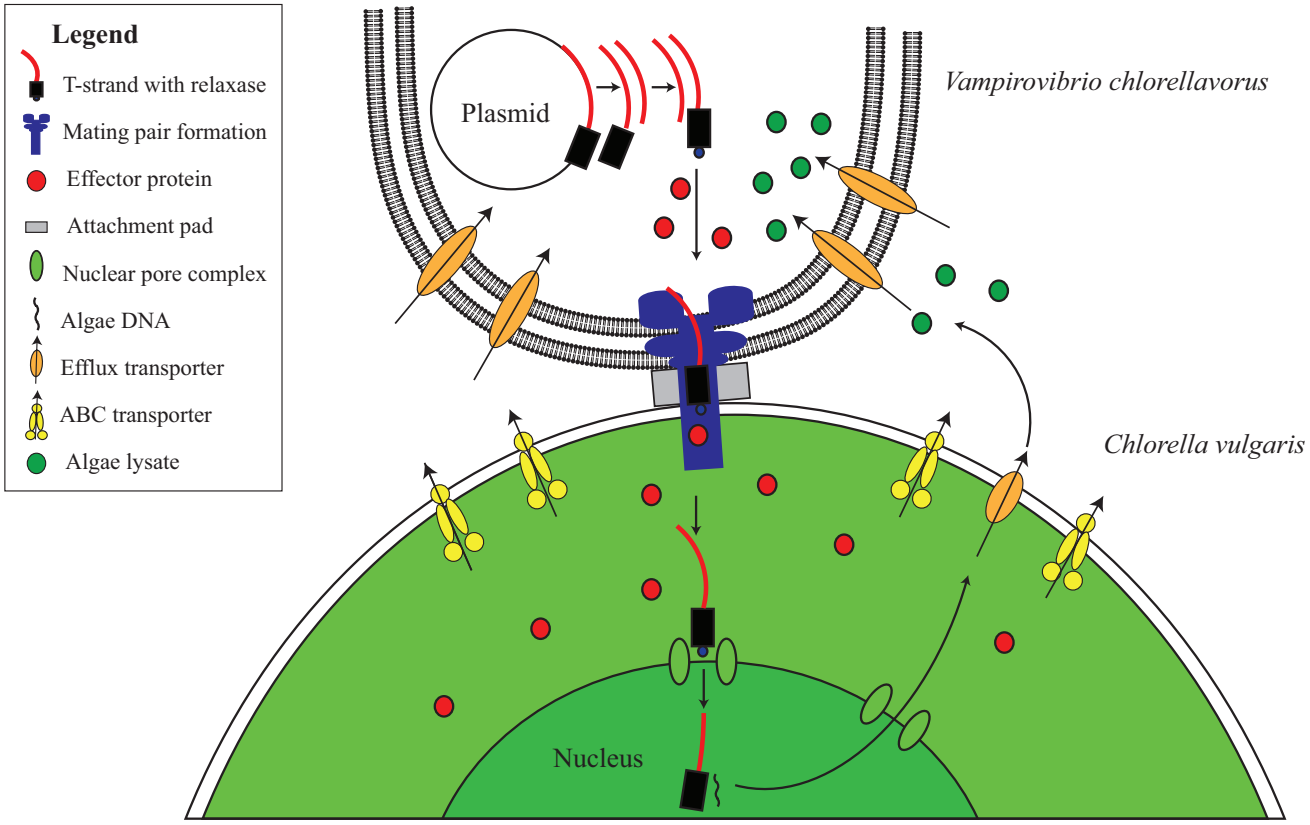
Ultrastructural studies have shown that *V. chlorellavorus* forms a discrete pad of unknown composition during attachment to *Chlorella* cells [8, 11]. Similar pads are involved in the attachment

of the uncultured predatory bacterium *Vampirococcus* to its bacterial prey, *Chromatium* [66]. Spikes of electron dense material have been observed to extend from the *V. chlorellavorus* pad into the *Chlorella* cell through the algal cell envelope [1]. We propose that the attachment pad and spike are a type IV secretion system (T4SS) fully encoded in the *V. chlorellavorus* genome in three operons (**Figure 3.2** and **Figure S3.4**). Phylogenetic analysis of the VirB4 ATPase (gene *trbE*), a highly conserved component of the T4SS used to classify these secretion systems [43] showed that the *V. chlorellavorus* orthologue is most closely related to a T-type conjugation system in *Nitrosomonas eutropha* (**Figure S3.5**). T-type conjugation T4SS are best known in *Agrobacterium tumefaciens* which form a secretion channel through which the T-strand (the strand destined for transfer) is passed into plant cells causing crown gall disease [67]. More generally, T-type conjugation systems can pass single stranded DNA and proteins into recipient cells [68]. Two of the T4SS operons of *V. chlorellavorus* are found on conjugative plasmids (**Figure S3.4**), which are predicted to be made singlestranded by their relaxases, nicking the DNA at the origin of transfer and transporting the T-strand to the *Chlorella* cell via the conjugation channel. The T-strand would then integrate into the *Chlorella* chromosome and be expressed [69] (**Figure 3.4**). Since the nature of the relationship between the two conjugating cells is predatory, we may expect that the T-strand would carry genes that facilitate ingestion of the *Chlorella* cell contents. No genes encoding hydrolytic enzymes were identified on the plasmids, though one encodes several efflux transporters (**Figure S3.4**; *see below*).

3.4.5.3 Phase iii: Ingestion

Five to seven days after *V. chlorellavorus* attachment, *Chlorella* cells remain intact but are devoid of cytoplasmic contents and contain only large vacuolated areas and membranous structures which are presumed to be organellar remains [1]. The *V. chlorellavorus* genome encodes numerous proteins that may be involved in the observed ingestion of *Chlorella* cell contents, including 108 proteases and 123 carbohydrate-active enzymes (**Tables S3.4** and **S3.5**). The majority of the latter group are glycoside hydrolases which are predicted to degrade polysaccharides and glycoproteins, major components of the *Chlorella* cell envelope [70] as well as starch and glycogen, which are diurnally stored as energy sources in *Chlorella* [71]. Extracellular proteases are produced by many bacterial pathogens and are commonly involved in the degradation of the host extracellular matrix, facilitating invasion and colonisation [72]. They have also been suggested as important factors in virulence for other predatory bacteria, for example *Bdellovibrio bacteriovorus* and *Micavibrio aeruginosa* [73, 74]. The *V. chlorellavorus* genome contains an alginate lyase, an enzyme that is able to degrade alginate via β-elimination cleavage of glycosidic bonds in the polysaccharide backbone [75]. Alginate is a common component of marine brown algae cell envelopes and intracellular material which is

Figure 3.4. Proposed conjugative mechanism



T4SS operons are found on two conjugative plasmids in *V. chlorellavorus*. The T-strands of the plasmids are predicted to be made single-stranded by plasmid-encoded relaxases (**Figure S3.4**), nicking the DNA at the origin of transfer and transporting the T-strands to the *C. vulgaris* cell via the mating pair formation. We predict that effector proteins (hydrolytic enzymes) synthesised in the bacterium are also transported via the mating pair formation. The T-strand enters the algal nucleus through a nuclear pore complex and is incorporated into a *C. vulgaris* chromosome. The effector proteins degrade the algae contents which are transported out of the algal cell via T-strand encoded transporters (**Figure S3.4**). The algal lysates are imported into the *V. chlorellavorus* cell providing energy and nutrients for replication.

targeted as a carbon and energy source by bacteria possessing alginate lyases [76]. *Chlorella* cells may similarly contain alginate supported by the finding of an alginate lyase gene in a *Chlorella* virus [77]. We propose that this suite of hydrolytic enzymes are synthesised in *V. chlorellavorus* and transported via the T4SS conjugation channel into the prey cell where they produce hydrolysates in the *Chlorella* cell (**Figure 3.3**). The T4SS plasmid-encoded efflux transporters (**Figure S3.4**) may facilitate the export of lysates from the *Chlorella* cell assuming that the T-strand is integrated and

expressed in *Chlorella* as is the case in *Agrobacterium* tumour formation [67]. Lysates exported into the surrounding milieu could then be imported into the attached *V. chlorellavorus* cell (and possibly neighbouring predatory cells) using a number of transport systems from the ATP-binding cassette (ABC) superfamily, the Major Facilitator Superfamily and/or permeases encoded in the bacterial genome (**Figure 3.2**). It is unlikely that *Chlorella* lysates would be directly transported into *V. chlorellavorus* cells via the conjugation channel as conjugation systems have only been shown to deliver protein or DNA substrates to eukaryotic target cells but not vice versa [69].

3.4.5.4 Phase iv: Binary fission

Attached *V. chlorellavorus* cells have been observed to divide by binary fission presumably using nutrients and energy derived from ingestion of *Chlorella* lysates, consistent with an obligate predatory lifestyle [1, 10]. The genome contains the cell division proteins required to replicate by this process, including the tubulin-like protein FtsZ, which is predicted to be highly expressed by PHX analysis, and the regulation of the placement of division site genes, *minC, -D and -E* [78].

3.4.5.5 Phase v: Release

A new lifecycle is started when progeny cells release from consumed *Chlorella* cells (**Figure 3.3**). Released cells then synthesise flagella to aid their dispersal and have a range of mechanisms to protect themselves from environmental stress as free-living organisms. The *V. chlorellavorus* genome encodes two superoxide dismutases, which convert O_2^- to H_2O_2 and O_2 [79] and one catalase-peroxidase, *katG*, a H_2O_2 scavenger [80]. Both of these enzymes can be used to combat oxidative stress that may be induced by environmental agents such as radiation or compounds that can generate intracellular O_2^- [79] or from the *Chlorella* [81]. The genome encodes a large and small conductance mechanosensitive channel protein that prevents cells from lysing upon sudden hypo-osmotic shock by releasing solutes and water [82]. It also encodes a protein containing a stress-induced bacterial acidophilic repeat motif and three copies of a universal stress protein (UspA), an autophosphorylating serine and threonine phosphoprotein [83]. In other stress conditions, such as temperature shock, starvation or the presence of oxidants or DNA-damaging agents, the expression of UspA is increased or decreased, which is known to be correlated with improved bacterial survival [84]. Beta-lactamases, cation/multidrug efflux pumps and ABC-type multidrug and solvent transport systems were identified (**Figure 3.2**) that could be used to eliminate antibiotics or toxins encountered in the environment [85, 86].

3.4.6 Comparison of *V. chlorellavorus* to other predatory bacteria

Pasternak et al. 2013, conducted a study of 11 predatory and 19 non-predatory bacterial genomes to define the 'predatome', the core gene set proposed for bacteria with predatory lifestyles [4]. The study found that the most striking difference between predators and non-predators is their method of synthesising isoprenoids. All predators, except for *M. aeruginosavorus*, encode the three essential enzymes used in the mevalonate pathway, which is uncommon in bacteria, whereas non-predators encode five essential enzymes for the more typical non-mevalonate pathway. It was suggested that predatory bacteria may have access to acetoacetyl-CoA pools in their prey cells, which is the first substrate used in the mevalonate pathway [4]. However, *V. chlorellavorus* lacks two of the three mevalonate pathway genes and instead encodes the non-mevalonate pathway (**Figure 3.2**). Twelve additional protein families were identified as specific to the predator set including those involved in chemotaxis, cell adhesion, degradation of polypeptides and benzoate, and four enzymes that may have evolved to scavenge essential metabolites [4]. *V. chlorellavorus* has orthologues of eight of these protein families and while lacking some of the specific adhesion and degradation genes (OrthoMCL OG4 39191, 26993, 21243, 18254), it encodes alternative proteins for these functions (see above). Eleven additional protein families were identified as specific to the non-predatory bacteria including those for riboflavin and amino acid synthesis, specifically tryptophan, phenylalanine, tyrosine, valine, leucine and isoleucine [4]. *V. chlorellavorus* has all but one of these "non-predatory" genes (OrthoMCL OG4 11203) which may reflect its phylogenetic novelty given that the core set analysis was based mostly on comparison of Proteobacteria [4]. We note that while *V. chlorellavorus* can make these particular compounds, its cofactor and amino acid biosynthesis repertoire is limited (5 cofactors, 15 amino acids).

3.4.7 Comparison of *V. chlorellavorus* to other Melainabacteria genomes

Consistent with all sequenced representatives of the class Melainabacteria [15, 16], *V. chlorellavorus* is missing all recognised photosynthesis genes including those for Photosystems I and II, chlorophyll and antennae proteins. This supports the hypothesis that photosynthetic cyanobacteria acquired photosystems after diverging from the ancestor of the Melainabacteria ([15, 16]; **Figure 3.1**). The *V. chlorellavorus* genome falls within the size range of previously reported Melainabacteria (1.8 to 5.5 Mbp) but has the highest GC content thus far (51.4%) compared with the GC content of other Melainabacteria who have a range of 27.5% to 49.4%. *V. chlorellavorus* is the second representative of the class inferred to be capable of oxidative phosphorylation as it contains a full respiratory chain (**Figure 3.2**), the other being *Obscuribacter phosphatis* [15]. *V. chlorellavorus* encodes a flagellum which is also found in some representatives of the order Gastranaerophilales (ACD20, MEL_B1 and MEL_B2). We inferred a phylogenetic tree for the conserved flagella marker gene, *fliI* [87] and

found that the Melainabacteria *fliI* genes form a monophyletic cluster consistent with their internal branching order in the genome tree (**Figure 3.1** and **Figure S3.6**) This association suggests that flagella were present in the cyanobacterial ancestor of the Gastranaerophilales and Vampirovibrionales and were subsequently lost at least once in the Gastranaerophilales (**Figure 3.1**). A global comparison of COG (clusters of orthologous groups) categories revealed that *V. chlorellavorus* has a functional distribution typical of other Melainabacteria genomes with the exception of genes involved in intracellular trafficking, secretion, and vesicular transport (**Figure S3.7**). *V. chlorellavorus* is overrepresented in this category due to a higher proportion of genes involved in Type IV secretion systems, which we posit to be important in the lifecycle of this predator (*see above*).

3.5 Conclusions

We have sequenced and assembled a near complete genome from a 36-year old lyophilised co-culture of the predatory bacterium *Vampirovibrio chlorellavorus*. Comparative gene and genome analyses confirm that *V. chlorellavorus* is a member of the Melainabacteria, a recently described non-photosynthetic class in the cyanobacterial phylum [15]. *V. chlorellavorus* is the first recognised member of the Cyanobacteria with a predatory lifecycle and we predict that it is the first predator to use a conjugative type IV secretion system similar to *Agrobacterium tumefaciens* to invade its host. It remains to be determined how widespread this phenotype is within the Melainabacteria and how it may have evolved from non-predatory cyanobacterial ancestors.

3.6 Acknowledgements

We thank Jim Prosser, Samantha Law and Tina Niven from NCIMB for their help with obtaining the co-cultures of *V. chlorellavorus* and *C. vulgaris* and Serene Lowe for preparing the DNA for sequencing and IMB, UQ for sequencing. We also thank Xuyen Le and Bryan Wee for discussions on motility and T4SS, Julien Guglielmini for data on T4SS, Rick Webb for inspection of S-layers in transmission electron microscopy images, Michael Nefedov for translation of Russian manuscripts and Nancy Lachner for attempts to extract RNA from the lyophilised cells.

3.7 References

1. Coder, D. and M. Starr, Antagonistic association of the chlorellavorus bacterium *("Bdellovibrio" chlorellavorus*) with *Chlorella vulgaris*. Current Microbiology, 1978. **1**(1): p. 59-64.
2. Stolp, H. and M.P. Starr, *Bdellovibrio bacteriovorus* gen. et sp. n., a predatory, ectoparasitic, and bacteriolytic microorganism. Antonie van Leeuwenhoek, 1963. **29**(1): p. 217-248.
3. Jurkevitch, E., Predatory behaviors in bacteria-diversity and transitions. Microbe-american society for microbiology, 2007. **2**(2): p. 67.
4. Pasternak, Z., et al., By their genes ye shall know them: genomic signatures of predatory bacteria. ISME J, 2013. **7**(4): p. 756-769.
5. Velicer, G.J., L. Kroos, and R.E. Lenski, Developmental cheating in the social bacterium *Myxococcus xanthus*. Nature, 2000. **404**(6778): p. 598-601.
6. Casida, L.E., Interaction of *Agromyces ramosus* with Other Bacteria in Soil. Applied and Environmental Microbiology, 1983. **46**(4): p. 881-888.
7. Saw, J.H., et al., Complete genome sequencing and analysis of *Saprospira grandis* str. Lewin, a predatory marine bacterium. Standards in Genomic Sciences, 2012. **6**(1): p. 84-93.
8. Gromov, B.V. and K.A. Mamkaeva, Electron microscopic study of parasitism by *Bdellovibrio chlorellavorus* bacteria on cells of the green alga *Chlorella vulgaris*. Tsitologiia, 1972. **14**(2): p. 256-60.
9. Coder, D.M. and L.J. Goff, The host range of the Chlorellavorous bacterium ("*Vampirovibrio chlorellavorus*"). Journal of Phycology, 1986. **22**(4): p. 543-546.
10. Gromov, B.V. and K.A. Mamkaeva, New genus of bacteria, *Vampirovibrio*, parasitizing chlorella and previously assigned to the genus *Bdellovibrio*. Mikrobiologiia, 1980. **49**(1): p. 165-7.
11. Mamkaeva, K.A. and O.V. Rybal'chenko, Ultrastructural characteristics of *Bdellovibrio chlorellavorus*. Mikrobiologiia, 1979. **48**(1): p. 159-61.
12. McDonald, D., et al., An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J, 2012. **6**(3): p. 610-8.
13. Quast, C., et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 2013. **41**(D1): p. D590-D596.
14. Corbett, L.L. and D.L. Parker, Viability of lyophilized cyanobacteria (blue-green algae). Applied and Environmental Microbiology, 1976. **32**(6): p. 777-80.
15. Soo, R.M., et al., An Expanded Genomic Representation of the Phylum Cyanobacteria. Genome Biology and Evolution, 2014. **6**(5): p. 1031-1045.

16. Di Rienzi, S.C., et al., The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. eLife, 2013. **2**: p. e01102.
17. Quail, M.A., H. Swerdlow, and D.J. Turner, Improved protocols for the illumina genome analyzer sequencing system. Current Protocols in Human Genetics, 2009. **62**: p. 18.2.1-18.2.27
18. Altschul, S.F., et al., Basic local alignment search tool. J Mol Biol, 1990. **215**(3): p. 403-10.
19. Seemann, T., Prokka: rapid prokaryotic genome annotation. Bioinformatics, 2014. **30**(14): p. 2068-2069.
20. Parks, D.H., et al., CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. PeerJ PrePrints, 2014. **2**: p. e554v1.
21. Langille, M.G. and F.S. Brinkman, IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics, 2009. **25**(5): p. 664-5.
22. Waack, S., et al., Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics, 2006. **7**(1): p. 142.
23. Markowitz, V.M., et al., IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics, 2009. **25**(17): p. 2271-2278.
24. Suzek, B.E., et al., UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 2007. **23**(10): p. 1282-1288.
25. Kanehisa, M., et al., The KEGG resource for deciphering the genome. Nucleic Acids Research, 2004. **32**: p. D277-D280.
26. Jones, P., et al., InterProScan 5: genome-scale protein function classification. Bioinformatics, 2014. **30**(9): p. 1236-1240.
27. Yin, Y., et al., dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Research, 2012. **40**: p. W445-51.
28. Rawlings, N.D., et al., MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Research, 2014. **42**: p. D503-D509.
29. Chaudhuri, R.R., et al., xBASE2: a comprehensive resource for comparative bacterial genomics. Nucleic Acids Research, 2008. **36**: p. D543-D546.
30. Bhaya, D., et al., Type IV pilus biogenesis and motility in the cyanobacterium *Synechocystis* sp. PCC6803. Molecular Microbiology, 2000. **37**(4): p. 941-951.
31. Karlin, S. and J. Mrázek, Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes. Journal of Bacteriology, 2000. **182**(18): p. 5238-5250.
32. Finn, R.D., et al., Pfam: the protein families database. Nucleic Acids Research, 2014. **42**: p. D222-30.

33. Haft, D.H., J.D. Selengut, and O. White, The TIGRFAMs database of protein families. Nucleic Acids Research, 2003. **31**(1): p. 371-3.
34. Markowitz, V.M., et al., IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Research, 2014. **42**: p. D568-D573.
35. Hyatt, D., et al., Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics, 2012. **28**(17): p. 2223-30.
36. Price, M.N., P.S. Dehal, and A.P. Arkin, FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution, 2009. **26**(7): p. 1641-1650.
37. Felsenstein, J., Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution, 1985. **39**(4): p. 783-791.
38. Ludwig, W., et al., ARB: a software environment for sequence data. Nucleic Acids Research, 2004. **32**(4): p. 1363-71.
39. Swofford, D.L. and J. Sullivan, Phylogeny inference based on parsimony and other methods using PAUP*. The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, 2009. Chapter 8, p. 267-312.
40. Stamatakis, A., RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 2006. **22**(21): p. 2688-2690.
41. Felsenstein, J., PHYLIP-phylogeny inference package (version 3.2). Cladistics, 1989. **5**: p. 163-166.
42. Guindon, S., et al., New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology, 2010. **59**(3): p. 307-321.
43. Guglielmini, J., F. de la Cruz, and E.P.C. Rocha, Evolution of Conjugation and Type IV Secretion Systems. Molecular Biology and Evolution, 2013. **30**(2): p. 315-331.
44. Chen, F., et al., OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research, 2006. **34**: p. D363-D368.
45. Hyatt, D., et al., Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 2010. **11**(1): p. 119.
46. Tatusov, R.L., et al., The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 2003. **4**: p. 41.
47. Parks, D.H., et al., STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics, 2014. **30**(21): p. 3123-4.

48. Shih, P.M., et al., Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(3): p. 1053-8.
49. White, C. L. and Gober, J. W. MreB: pilot or passenger of cell wall synthesis. Trends in Microbiology, 2012. **20**(2):p. 74-79.
50. Varma, A. and K.D. Young, In *Escherichia coli*, MreB and FtsZ Direct the Synthesis of Lateral Cell Wall via Independent Pathways That Require PBP 2. Journal of Bacteriology, 2009. **191**(11): p. 3526-3533.
51. Beveridge, T.J., Structures of Gram-Negative Cell Walls and Their Derived Membrane Vesicles. Journal of Bacteriology, 1999. **181**(16): p. 4725-4733.
52. Sára, M. and U.B. Sleytr, S-Layer Proteins. Journal of Bacteriology, 2000. **182**(4): p. 859-868.
53. Šmarda, J., et al., S-layers on cell walls of cyanobacteria. Micron, 2002. **33**(3): p. 257-277.
54. Preisig, O., et al., A high-affinity *cbb*$_3$-type cytochrome oxidase terminates the symbiosis-specific respiratory chain of *Bradyrhizobium japonicum*. Journal of Bacteriology, 1996. **178**(6): p. 1532-8.
55. Wadhams, G.H. and J.P. Armitage, Making sense of it all: bacterial chemotaxis. Nature Reviews Molecular Cell Biology, 2004. **5**(12): p. 1024-1037.
56. Macnab, R.M., How bacteria assemble flagella. Annual Review of Microbiology, 2003. **57**: p. 77-100.
57. Duggan, P.S., P. Gottardello, and D.G. Adams, Molecular Analysis of Genes in *Nostoc punctiforme* Involved in Pilus Biogenesis and Plant Infection. Journal of Bacteriology, 2007. **189**(12): p. 4547-4551.
58. Freitas, T.A.K., S. Hou, and M. Alam, The diversity of globin-coupled sensors. FEBS Letters. **552**(2): p. 99-104.
59. Bhoo, S.-H., et al., Bacteriophytochromes are photochromic histidine kinases using a biliverdin chromophore. Nature, 2001. **414**(6865): p. 776-779.
60. Chamot-Rooke, J., et al., Posttranslational Modification of Pili upon Cell Contact Triggers N. meningitidis Dissemination. Science, 2011. **331**(6018): p. 778-782.
61. Pizarro-Cerdá, J. and P. Cossart, Bacterial Adhesion and Entry into Host Cells. Cell, 2006. **124**(4): p. 715-727.
62. Shin, S., et al., *Escherichia coli* outer membrane protein A adheres to human brain microvascular endothelial cells. Biochemical and Biophysical Research Communications, 2005. **330**(4): p. 1199-204.

63. Smani, Y., M.J. McConnell, and J. Pachón, Role of Fibronectin in the Adhesion of *Acinetobacter baumannii* to Host Cells. PLoS ONE, 2012. **7**(4): p. e33073.
64. Whittaker, C.A. and R.O. Hynes, Distribution and Evolution of von Willebrand/Integrin A Domains: Widely Dispersed Domains with Roles in Cell Adhesion and Elsewhere. Molecular Biology of the Cell, 2002. **13**(10): p. 3369-3387.
65. Nielsen, H.V., et al., The Metal Ion-Dependent Adhesion Site Motif of the *Enterococcus faecalis* EbpA Pilin Mediates Pilus Function in Catheter-Associated Urinary Tract Infection. mBio, 2012. **3**(4): p. e00177-12.
66. Guerrero, R., et al., Predatory prokaryotes: Predation and primary consumption evolved in bacteria. Proceedings of the National Academy of Sciences of the United States of America, 1986. **83**(7): p. 2138-2142.
67. Christie, P.J., Type IV secretion: the Agrobacterium VirB/D4 and related conjugation systems. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 2004. **1694**(1–3): p. 219-234.
68. Alvarez-Martinez, C.E. and P.J. Christie, Biological diversity of prokaryotic type IV secretion systems. Microbiology and Molecular Biology Reviews, 2009. **73**(4): p. 775-808.
69. Cascales, E. and P.J. Christie, The versatile bacterial type IV secretion systems. Nature Reviews Microbiology, 2003. **1**(2): p. 137-149.
70. Gerken, H.G., B. Donohoe, and E.P. Knoshaug, Enzymatic cell wall degradation of *Chlorella vulgaris* and other microalgae for biofuels production. Planta, 2013. **237**(1): p. 239-53.
71. Nakamura, Y. and S. Miyachi, Effect of Temperature on Starch Degradation in *Chlorella vulgaris* 11h Cells. Plant and Cell Physiology, 1982. **23**(2): p. 333-341.
72. Kennan, R.M., et al., The Subtilisin-Like Protease AprV2 Is Required for Virulence and Uses a Novel Disulphide-Tethered Exosite to Bind Substrates. PLoS Pathogens, 2010. **6**(11): p. e1001210.
73. Rendulic, S., et al., A Predator Unmasked: Life Cycle of *Bdellovibrio bacteriovorus* from a Genomic Perspective. Science, 2004. **303**(5658): p. 689-692.
74. Wang, Z., D. Kadouri, and M. Wu, Genomic insights into an obligate epibiotic bacterial predator: *Micavibrio aeruginosavorus* ARL-13. BMC Genomics, 2011. **12**(1): p. 453.
75. Lamppa, J.W., et al., Genetically Engineered Alginate Lyase-PEG Conjugates Exhibit Enhanced Catalytic Function and Reduced Immunoreactivity. PLoS ONE, 2011. **6**(2): p. e17042.

76. Wong, T.Y., L.A. Preston, and N.L. Schiller, Alginate lyase: Review of Major Sources and Enzyme Characteristics, Structure-Function Analysis, Biological Roles, and Applications. Annual Review of Microbiology, 2000. **54**(1): p. 289-340.
77. Suda, K., et al., Evidence for a novel Chlorella virus-encoded alginate lyase. FEMS Microbiology Letters, 1999. **180**(1): p. 45-53.
78. Lutkenhaus, J. and S.G. Addinall, Bacterial cell division and the Z ring. Annual Review of Biochemistry, 1997. **66**: p. 93-116.
79. Cabiscol, E., J. Tamarit, and J. Ros, Oxidative stress in bacteria and protein damage by reactive oxygen species. International Microbiology, 2000. **3**(1): p. 3-8.
80. Jittawuttipoka, T., et al., The Catalase-Peroxidase KatG Is Required for Virulence of *Xanthomonas campestris* pv. campestris in a Host Plant by Providing Protection against Low Levels of H2O2. Journal of Bacteriology, 2009. **191**(23): p. 7372-7377.
81. Mallick, N. and F.H. Mohn, Reactive oxygen species: response of algal cells. Journal of Plant Physiology, 2000. **157**(2): p. 183-193.
82. Birkner, J.P., B. Poolman, and A. Koçer, Hydrophobic gating of mechanosensitive channel of large conductance evidenced by single-subunit resolution. Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(32): p. 12944-12949.
83. Kvint, K., et al., The bacterial universal stress protein: function and regulation. Current Opinion in Microbiology, 2003. **6**(2): p. 140-145.
84. Jenkins, R., N. Burton, and R. Cooper, Effect of manuka honey on the expression of universal stress protein A in meticillin-resistant *Staphylococcus aureus*. International Journal of Antimicrobial Agents, 2011. **37**(4): p. 373-376.
85. Frère, J.-M., Beta-lactamases and bacterial resistance to antibiotics. Molecular Microbiology, 1995. **16**(3): p. 385-395.
86. Lubelski, J., W.N. Konings, and A.J.M. Driessen, Distribution and Physiology of ABC-Type Transporters Contributing to Multidrug Resistance in Bacteria. Microbiology and Molecular Biology Reviews, 2007. **71**(3): p. 463-476.
87. Minamino, T. and K. Namba, Distinct roles of the FliI ATPase and proton motive force in bacterial flagellar protein export. Nature, 2008. **451**(7177): p. 485-488.

Chapter 4: Population genomics and transcriptomics of two Obscuribacterales populations recovered from palsa in Stordalen Mire, northern Sweden

4.1 Abstract

Recently, a class of non-photosynthetic Cyanobacteria, called the Melainabacteria, have been recovered from a range of habitats including mammalian guts, an aquifer, bioreactor and freshwater. In this study, metagenomic sequencing from a permafrost sample recovered from palsa in Stordalen mire, northern Sweden, led to the recovery of two near-complete population genomes from the Melainabacteria. Genome annotation of these two populations revealed that they may have strategies for adapting to the cold environment of the palsa, including encoding for chaperones and stress proteins, genes for carbon and energy reserves, cryoprotectants, oxidative stress and cell membrane adaptations.

4.2 Introduction

Permafrost, ground that stays frozen for at least 2 years [1] comprises ~24% of the land mass in the northern hemisphere [2]. Permafrost areas can be characterised as continuous (a continuous sheet of frozen material comprising of more than 80% of the ground underlain by permafrost), discontinuous (areas of frozen material and areas that melt in the summer) or sporadic (an environment where less than 30% of the ground is permafrost). Continuous permafrost is found in the Arctic and Polar zones, whereas discontinuous is found in the Subarctic zone, and can be comprised of palsa (a peaty permafrost mound), bog (poorly-drained lowland underlain by organic-rich sediments) and fen (a peatland dominated by vegetation and slight depressions which contain shallow pools) [1].

The layer found above the permafrost, the active layer, thaws in the summer and refreezes in winter posing a challenge for resident microbes due to low temperatures, increased exposure to UV radiation and cycles of freeze-thaw [3]. In order to survive these harsh environments, microbes have to develop adaptive mechanisms. Most of what is known about microbial metabolism in permafrost comes from enrichment studies and includes adaptations such as cold shock proteins and cold acclimation proteins [4-6], sporulation [7, 8] and cryoprotectants to prevent cell-lysis during freeze-thaw cycles [9, 10].

Microscopic counts indicate microbial numbers of 10^7-10^9 cells g^{-1} in most permafrost environments, however viable counts account for only 0.001-10% of the observed cells [3]. Culture-independent surveys targeting the 16S rRNA gene in permafrost samples shows that bacterial diversity is greater than first expected in these harsh environments, with the discovery of microorganisms not found previously in culture from permafrost [11]. Moreover, metagenomics studies are revealing biochemical pathways that may be present or potentially expressed by the microbial communities without the biases of cultivation [12].

Amongst the bacterial phyla found in permafrost as a minority group is the Cyanobacteria. Vishnivetskaya and colleagues (2001) cultured thirty viable non-axenic cyanobacterial strains from 28 Siberian permafrost cores. These included members of the filamentous heterocystous and non-heterocystous cyanobacteria. They also attempted to culture Cyanobacteria from Antarctic permafrost but were unsuccessful. However, Blanco and colleagues (2009) sequenced the 16S rRNA gene from genomic DNA recovered from permafrost at Deception Island, Antarctica and found nitrogen-fixing cyanobacteria at both the surface and 160-200 cm below the surface.

A non-photosynthetic class of Cyanobacteria, the Melainabacteria, have been recovered and described from an aquifer [15], human gut [15, 16], koala gut, bioreactors [16] and freshwater [17]. Six orders have been identified within the class Melainabacteria and four have been assigned names based on habitat and metabolic potential; Gastranaerophilales (mammal faeces), Caenarcaniphilales (anaerobic bioreactor), Obscuribacterales (bioreactor) and Vampirovibrionales (*V. chlorellavorus*). In this study we recovered two population genomes belonging to the order Obscuribacterales from a palsa metagenome sample collected from Stordalen Mire, Sweden in summer (July) 2012 [18]. Potential adaptations that may allow these organisms to survive in the cold palsa environment were identified by comparative genomics and transcriptomics was used to determine which genes were being expressed *in situ*.

4.3 Methods and Materials

4.3.1 Sample collection

Sample P3D was collected from the active layer palsa (30-33cm below the surface) in Stordalen Mire, northern Sweden in July 2012 (**Figure S4.1**). A corer was used to collect the sample and the temperature was recorded (0.8°C). No ice was present, however samples were moist and composed of organic material. The cores were sectioned and 0.5g from the bottom core was placed in a sterile cryotube containing ~3 volumes of LifeGuard soil preservation solution (MoBio Laboratories, Carlsbad, CA, USA) and stored at -80°C until further processing.

4.3.2 DNA and RNA extraction and sequencing

The DNA and RNA from sample P3D was extracted as in Mondav et al., 2014. Briefly, the DNA and RNA were extracted using a MoBio PowerMax Nucleic Acid Extraction kit (using a bead-beater) and according to protocol (MoBio Laboratories, Carlsbad, CA, USA) followed by an ethanol precipitation. RNA was recovered by DNase digestion of DNA and amplified by PCR. DNA was recovered by RNase digestion. A phenol:chloroform:isoamyl alcohol cleanup was used to remove enzymes followed by an additional ethanol precipitation.

For metagenomic sequencing, ~200 ng of genomic DNA (gDNA) was used to construct two paired-end libraries. The libraries were prepared using an Illumina TruSeq Nano DNA Sample Preparation kit (Illumina, CA, USA) according to protocol. The first library was sequenced on one lane of an Illumina HiSeq 2000 (2 x 100 bp) using a TruSeq SBS kit v3.HS (Illumina, CA, USA) at the Institute of Molecular Biosciences, University of Queensland (IMB, UQ). The second library was sequenced on one lane of an Illumina NextSeq (2 x 150 bp) using a NextSeq 500 v2 kit (Illumina, CA, USA) at the Australian Centre for Ecogenomics, University of Queensland (ACE, UQ). Paired-end sequencing reads from the HiSeq and NextSeq were processed using Seqprep (https://github.com/jstjohn/SeqPrep) and Nesoni (https://github.com/Victorian-Bioinformatics-Consortium/nesoni) to remove adaptors and merge paired reads with overlaps into single reads. The trimmed reads were then assembled using CLC MainWorkbench (v.7.0) (CLC bio, Aarhus, Denmark) with a kmer size of 63.

4.3.3 Determining relative abundance and binning the Melainabacteria genomes

GraftM (https://github.com/geronimp/graftM) version r2439db was used to determine the relative abundance of Melainabacteria in the P3D sample using the trimmed metagenomic paired-end sequencing reads and the Greengenes 2013 97% OTUs (operational taxonomic units) as a reference [19].

The assembled contigs from P3D were queried against a BLAST database constructed from a dereplicated set of genes from all finished genomes from the Integrated Microbial Genomes (IMG) database (v.4.1) [20] and all genes from the Melainabacteria genomes [15, 16] using BLASTP (v.2.2.30+) with default parameters. Contigs containing open reading frames (ORFs) identified as the best hit to the Melainabacteria [15, 16] were extracted from the metagenomic data.

Sequencing reads from the palsa sample were mapped against the extracted contigs using BamM (http://minillinim.github.io/BamM/) to obtain coverages to identify potential population genomes.

GC content for each contig was obtained using seqFilter.pl (https://github.com/ctSkennerton/scriptShed/blob/master/seqFilter.pl). Two Melainabacteria population genomes were identified from the coverage and GC information and the contigs belonging to these two groups were extracted with the exclusion of contigs < 2kbp. Contigs with > 10% of total ORFs (as identified by the BLASTP results) belonging to Melainabacteria were kept and contigs with < 10% were discarded. CheckM (v.0.9.7) was used to identify the completeness and contamination of the population bins based on the presence and absence of 104 conservative bacterial marker genes [21].

4.3.4 Phylogenetic tree

A concatenated gene tree was inferred in order to establish the phylogenetic relationship of the two Melainabacteria genomes from P3D. A set of 83 marker genes [16] was used to align the two population genomes with 2,311 complete bacterial and archaeal reference genomes obtained from IMG (v.4.0) [20] and 12 Melainabacteria population genomes [15-17] using HMMER3 [22]. Individual gene trees were constructed using FastTree v.2.1.7 with default parameters for bootstrapping [23]. CheckM (v.0.9.7) [21] was used to calculate the amino acid identity (AAI) between the two population genomes from P3D and *Ca.* O. phosphatis [16].

4.3.5 Genome annotation

The two draft population genomes from P3D were submitted to the Joint Genome Institute Integrated Microbial Genomes Expert Review (JGI IMG-ER) for automated gene annotation [24]. Both genomes are deposited in the JGI IMG-ER database under the accession numbers 2603880200 (P3DObs1) and 2603880201 (P3DObs2). Additionally, the population genomes were annotated using PROKKA (v.1.8) [25] with default parameters. A partial 16S rRNA gene (347 bp) was identified in P3DObs2 and a BLAST search was used to identify the closest neighbour from the Greengenes October 2011 database [19]. The tRNAscan-SE (v.1.21) Search Server was used to identify tRNA genes in the genomes [26]. KEGG maps and gene annotations were used to reconstruct the metabolism of the two Obscuribacterales genomes. Carbohydrate-active enzymes were annotated using dbCAN HMMs and HMMER3 (hmmscan) with default parameters [27] and amino acid usage was calculated using CompareM (v.0.0.2) (https://github.com/dparks1134/CompreM). The CRISPRfinder programme was used to identify potential Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and CRISPR-associated (cas) genes [28]. Detection of genes encoding for natural products were identified using antibiotics and Secondary Metabolite Analysis SHell (antiSMASH) (v.3.0.0) [29] and the bacteriocins were compared to those in BACTIBASE [30] using BLASTP (v.2.2.29+). KEGG

maps, the MetaCyc database [31] and the gene annotations from JGI IMG-ER were used to reconstruct the metabolism of the two population genomes and a composite metabolic cartoon was prepared in Adobe Illustrator CS6 (Adobe, CA, USA). The COG categories were calculated using an in-house script, cogTable.py, which performs a homology search against the COG database using uclust and an e-value cut-off of 0.01, a 30% identity cut-off and an alignment length of at least 70%.

4.3.6 Metatranscriptomics

Ribosomal RNA (rRNA) removal, complementary DNA (cDNA) synthesis and cDNA library preparation was prepared with a ScriptSeq Complete Kit (Bacteria) – Low input (Illumina, CA, USA). All amplified cDNA was sequenced on 1/8th of a lane of NextSeq at ACE, UQ using a NextSeq 500 v2 kit (Illumina, CA, USA). The transcriptomics paired-end reads had adaptors removed and were merged with overlaps into single reads using Seqprep (https://github.com/jstjohn/SeqPrep) and Nesoni (https://github.com/Victorian-Bioinformatics-Consortium/nesoni). Residual rRNA was removed using SortMeRNA (v.1.9) [32] filtered by comparison against the Silva OTU database (v.111) [33]. The transcripts for P3D were mapped against a concatenated fasta file of both Obscuribacterales genomes using BamM (http://minillinim.github.io/BamM/). Fragments mapped per kilobase of gene length (FPKG) was calculated for each gene using sam2fpkg (https://github.com/minillinim/sam2fpkg) and the ORFs that had an FPKG value of < 0 were identified as being up-regulated in the population genomes. Transcriptomics data was visualised in Tablet [34] to check that sequence reads were unidirectional.

4.4 Results and Discussion

4.4.1 Metagenome and metatranscriptome data summary

A total of 86.83 Gbp of metagenomic and 9.03 Gbp of metatranscriptomic sequencing data was obtained from the P3D palsa sample. The metagenomic paired sequence reads were assembled into 3,332,579 contigs ≥ 2 kb in length with a N50 of 2,884.

4.4.2 Obscuribacterales population genomes and gene expression

The GraftM results showed that 0.9% of the sequencing reads from P3D mapped to Obsucribacterale*s* 16S rRNA genes from the Greengenes database [19]. From the BLASTP results, 15,473 open reading frames (ORFs) had top hits to *Ca.* O. phosphatis and two Obscuribacterales genome populations were identified based on coverage and GC content (P3DObs1 and P3DObs2). After filtering contigs based on size (removing contigs < 2kbp) and only including contigs that have >10% of total ORFs with a top hit to *Ca.* O. phosphatis, P3DObs1 had a read coverage of 55 to 63-

fold and a GC range of 46-54% and P3DObs2 had a read coverage of 117 to 133-fold and a GC of 46-54% **(Table 4.1)**. P3DObs1 consists of 7.03 Mbp in 187 contigs with an average GC content of 48.9% and is estimated to be 88.9% complete and 2.6% contaminated (two copies of PF06071 (Protein of unknown function (DUF933)) and PF00318 (ribosomal protein S2)). The genome is predicted to encode 6,570 open reading frames (ORFs), with 53.5% of the ORFs having a predicted function. In addition, a partial 23S rRNA (347 bp) and an unusually large number of tRNA genes (77) for the 20 regular amino acids as well as selenocysteine were identified. It has been suggested that a large number of tRNA genes may participate in adaptation to cold conditions, allowing for faster growth as lower temperatures may limit the speed of transcription and translation [35]. P3DObs2 consists of 6.87 Mbp in 139 contigs with an average GC content of 48.6% and is estimated to be 86.5% complete and 2.6% contaminated (two copies of PF00318 (ribosomal protein S2) and PF01121 (Dephospho-CoA kinase) (**Table 4.1**). The genome is predicted to code for 6,119 ORFs with 57.9% of the ORFs having a predicted function. A partial 5S rRNA gene (116 bp), a partial 16S rRNA gene (345 bp) and 46 tRNA genes for all the regular amino acids except asparagine, were identified. Both P3DObs genomes are larger than previously sequenced Melainabacteria (1.8– 5.5 Mbp) (**Table 4.1**) and contain a number of genes associated with mobile genetic elements (0.6% of total genes for both) encoding phage integrases, transposases and other phage elements. In addition, P3DObs1 contains a potential Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) (347 bp) however no CRISPR-associated (cas) genes were identified.

Mapping of metatranscriptomic reads to the two Obscuribacterales population genomes indicated that 10.9% and 16.7% of predicted genes in P3DObs1 and 2 were expressed in the palsa habitat. Expression levels ranged from zero to 4,424.6 FPKG (fragments mapped per kilobase of gene length) for P3DObs1 with an average of 81 (excluding non-expressed genes) and zero to 3,677.2 FPKG for P3DObs1 with an average of 121. A large portion of the transcripts from P3DObs1 and 2 mapped to hypothetical proteins and tRNAs, with only 50% and 54% respectively, having a known function.

4.4.3 Phylogeny and taxonomy

A phylogenetic tree based on the concatenation of 83 conserved single copy marker genes [16] using 211 genomes representing 22 bacterial phyla and representatives from the archaeal domain, produced a robust topology placing P3DObs1 and 2 in the class Melainabacteria, monophyletic with the order Obscuribacterales (**Figure 4.1**). In addition, genomic diversity of the population genomes was further assessed using average amino acid identity (AAI). The two

Table 4.1. Genome statistics for Melainbacteria representatives

Population genome	Order	# of contigs	Estimated genome size (Mbp)	GC	# of genes[a]	rRNAs	tRNAs	Estimated complete-ness (%)[b]	Estimated contamin-ation (%)[b]	Study
P3DObs1	Obscuribacterales	187	7.0	48.9	6,652 (6,570)	Partial 23S	77	88.9%	2.6%	Present study
P3DObs2	Obscuribacterales	139	6.9	48.5	6,169 (6,119)	Partial 5S and 16S	46	86.5%	2.6%	Present study
Ca. Obscuribacter phosphatis	Obscuribacterales	8	5.5	49.4	4,392 (4,342)	5S,16S,23S	44	98.3% [c]	2.7% [c]	[16]
MH_37	Gastranaerophilales	157	2.2	34.1	2,402 (2,360)	-	36	100.0%	1.7%	[16]
Zag_1	Gastranaerophilales	322	2.0	34.9	2,194 (2,160)	-	31	94.8%	0.0%	[16]
Zag_111	Gastranaerophilales	65	2.2	36.7	2,313 (2,257)	5S, 16S, 23S	45	96.6%	3.6%	[16]
Ca. Gastranaerophilus phascolarctosicola	Gastranaerophilales	14	1.8	38.5	1,838 (1,799)	16S	37	100.0%	1.7%	[16]
MEL_A1	Gastranaerophilales	1	1.9	33.0	1,879 (1,832)	5S, 16S, 23S	40	100.0%	0.0%	[15]
MEL_B1	Gastranaerophilales	21	2.3	35.4	2,269 (2,219)	5S, 16S, 23S	42	98.3%	0.0%	[15]
MEL_B2	Gastranaerophilales	26	2.3	36.3	2,262 (2,215)	5S, 16S, 23S	41	100.0%	0.0%	[15]
MEL_C1	Gastranaerophilales	4	2.1	34.1	2,162 (2,120)	5S, 16S, 23S	36	100.0%	0.0%	[15]
ACD20	Gastranaerophilales	185	2.7	33.5	2,455 (2,325)	5S, 23S	37	100.0%	3.5%	[15]
Ca. Caenarcanum bioreactoricola	Caenarcaniphilales	67	1.8	27.5	1,917 (1,870)	16S, 23S	41	93.1%	1.7%	[16]

Vampirovibrio chlorellavorus	Vampirovibrionales	26 + 2[d]	3.0	51.4	2,798 (2,844)	16S, 23S	41	100.0%	1.7%	[17]

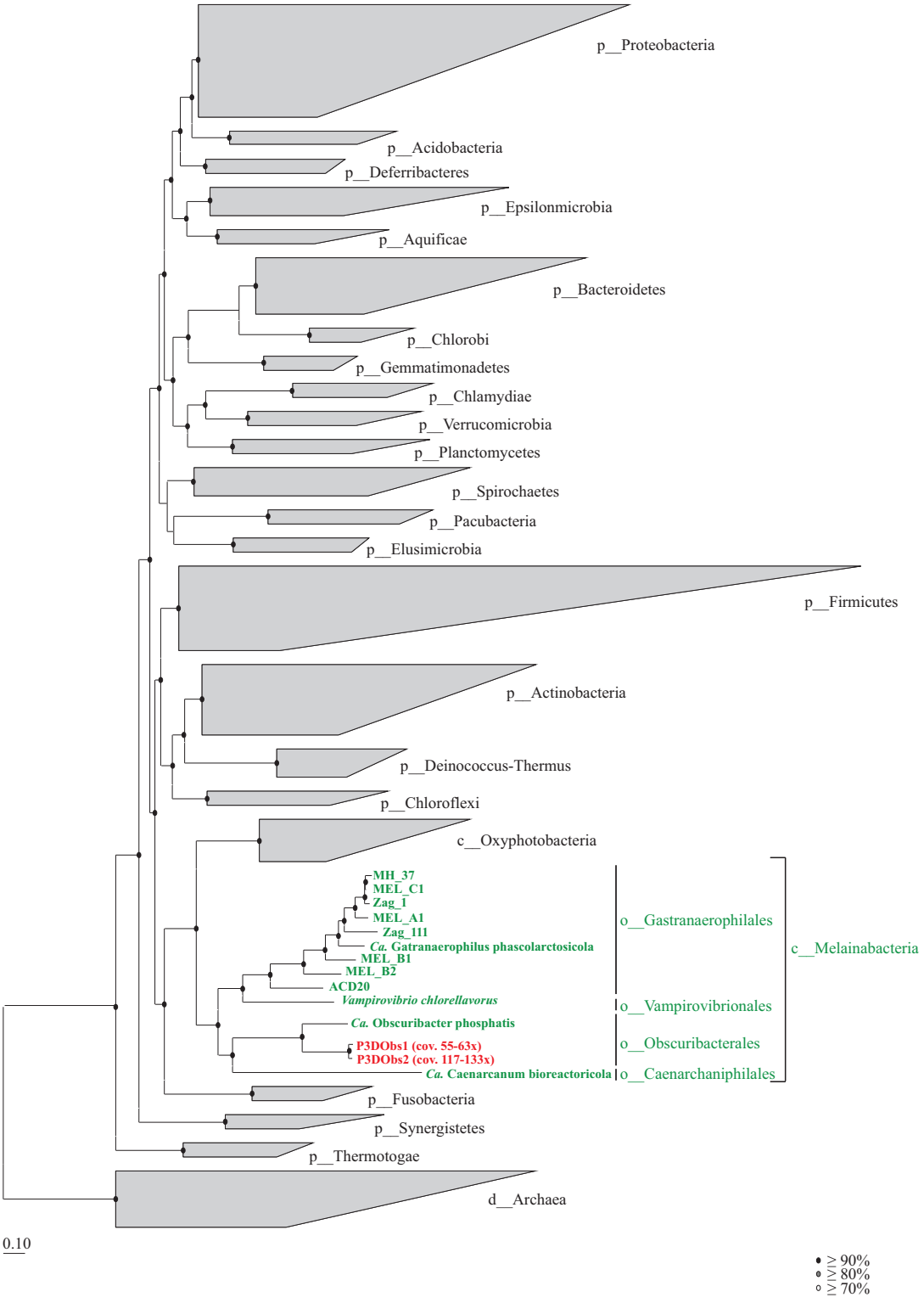
a. The numbers in brackets represent the number of protein coding genes

b. Estimated completeness and contamination is based on the presence and absence of 104 bacterial conserved single copy markers

c. Estimated completeness and contamination was predicted with checkM version 0.9.7 [21]

d. *Vampirovibrio chlorellavorus* contains 2 circular plasmids

Figure 4.1. Maximum likelihood concatenated gene tree using 83 single copy marker genes



Class Melainabacteria genomes are in green with the two Obscuribacterale*s* genomes recovered from P3D in red. Class Oxyphotobacteria and Melainabacteria belong to the phylum Cyanobacteria. p__ represents phylum, c__ represents class and o__ represents order. Black circles in the tree represents nodes with >90% bootstrap support, grey circles represent nodes with >80% bootstrap support and white circles in the tree represent nodes with >70% bootstrap support.

Obscuribacterales genomes from P3D were closely related (89.1% AAI; 3,774 orthologous genes), likely representing two species of the same genus. The P3DObs genomes were more distantly related to *Ca.* O. phosphatis sharing an AAI of 55.4% (1,799 orthologous genes) with P3DObs1 and 55.3% (1,923 orthologous genes) with P3DObs2, suggesting that they belong to the same family [36]. A partial 16S rRNA gene (345 bp) recovered from the P3DObs2 genome which was 90% identical to its ortholog in *Ca.* O. phosphatis, further indicated that the P3DObs genomes are members of the same family [36].

4.4.4 Cell wall and shape

Consistent with the structure of the cell wall found in other Melainabacteria and the Oxyphotobacteria [16], the two P3DObs genomes have a complement of genes indicative of a Gram-negative (diderm) cell envelope composed of Lipid A, O-antigens and lipopolysaccharides (LPS) [37] (**Table S4.1**). In addition, they also contain the proteins MreBCD, penicillin-binding protein 2 (MrdA) and either RodA or RodZ, which are responsible for the elongation and septation of the peptidoglycan layer resulting in a rod-shaped cell [38]. The transcripts from P3DObs1 and 2 mapped to three genes; O-antigen ligase, a phosphorylase involved in LPS biosynthesis and an acetyltransferase used for lipid metabolism, at very low levels (**Table S4.1**), however, transcripts for cell shape were absent from both genomes suggesting that at the time of sampling, the cells were not actively dividing.

4.4.5 Metabolism of Obscuribacterales genomes

4.4.5.1 Energy metabolism

Similar to *Ca.* O. phosphatis, both permafrost Obscuribacterales are facultative anaerobes capable of respiration and fermentative metabolism. For aerobic growth, the genomes have an incomplete tricarboxylic (TCA) cycle but like *Ca.* O. phoshatis, P3DObs2 contains the genes required for the glyoxylate shunt (malate synthase and isocitrate lyase) allowing it to bypass the two decarboxylation steps of the TCA cycle [39] whereas the glyoxylate shunt is absent in P3DObs1 (**Figures 4.2** and **4.3**). Both genomes have an electron transport chain consisting of complexes I, III, IV and an F-type ATPase, and high-affinity terminal oxidases, cytochrome cbb_3 and cytochrome bd, allowing these bacteria to live in microaerophilic environments [40]. However, of the entire electron transport chain, only subunit one from the cytochrome bd terminal oxidase in P3DObs1 was transcriptionally expressed (**Figure 4.2**) suggesting that the bacteria were not actively respiring at the time of sampling. In addition to aerobic respiration, both genomes have the potential to respire anaerobically using either fumarate (P3DObs1) or nitrate (P3DObs2) as electron acceptors. One of two subunits (*sdhA*) from succinate dehydrogenase/fumarate reductase was expressed at low

levels in P3DObs1 and three of the four subunits of nitrate reductase (*narGHI*) was highly expressed in P3DObs2, suggesting that both are performing anaerobic respiration (**Figures 4.2** and **4.3**). P3DObs1 and 2 have the potential to degrade ethanol, as they both encode aldehyde dehydrogenase and alcohol dehydrogenase.

The P3DObs1 genome encodes a Ni,Fe-hydrogenase I (membrane-bound uptake hydrogenase) which may link the oxidation of H_2 to the reduction of anaerobic electron acceptor fumarate (anaerobic respiration) or to O_2 (aerobic respiration) with the recovery of energy in the form of a proton motive force [42]. Both genomes encode formate hydrogenlyase (Ni,Fe-hydrogenase III; small and large subunits, *hycCDEF*) leading to the formation of H_2 and CO_2 from formate during fermentative growth. The genes required for assembly of the NiFe hydrogenases (*hoxEFUYH* and *hypABCDEF*) [43, 44] are also encoded on both genomes. Only two NiFe hydrogenase genes were expressed in P3DObs2 but at very low levels.

Both genomes also contain ethanolamine ammonia lyase light and heavy chains, providing the potential of using ethanolamine as a carbon and nitrogen source in the form of acetaldehyde and ammonia [41].

4.4.5.2 Carbohydrate metabolism

Both genomes have multiple enzymes for degrading monosaccharides (glucose, fructose, xylose and galactose) and polysaccharides (starch, glycogen and cellulose) (**Figures 4.2** and **4.3**). These include alpha-amylase, alpha and beta-glucosidases, cellulase and cellobiase (**Table S4.2**). P3DObs2 had a very high transcription level for cellobiase, which hydrolyses the cellobiose disaccharide into glucose and both genomes have high transcription levels for beta-galactosidase, which can break down the products further for entry into the Embden-Meyerhof-Parnas (EMP) pathway. In addition, P3DObs2 is able to convert the monosaccharide mannose using mannose-6-phosphate isomerase. Both P3DObs genomes encode the genes required for the upper phase of the EMP pathway but are missing half of the lower phase genes. This may be due to the genomes being incomplete. Both genomes encode all genes required for the non-oxidative branch of the pentose phosphate pathway, which can be used as a source of carbon skeletons for the synthesis of nucleotides, aromatic amino acids, pheylpropanoids and their derivatives [45], however, both are missing the enzymes for the oxidative branch, which is used to produce NADPH.

Figure 4.2. Metabolic reconstruction of P3DObs1

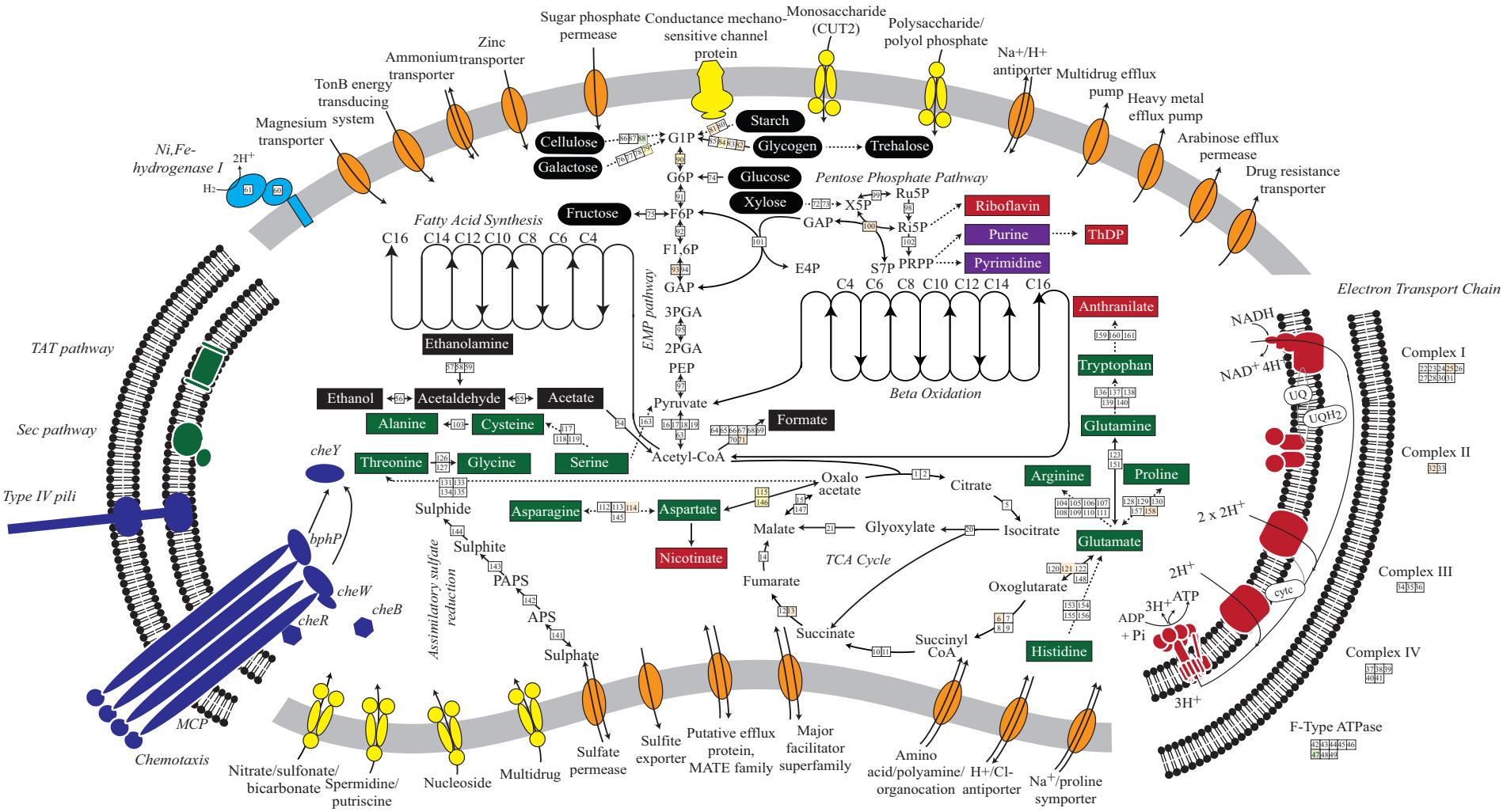
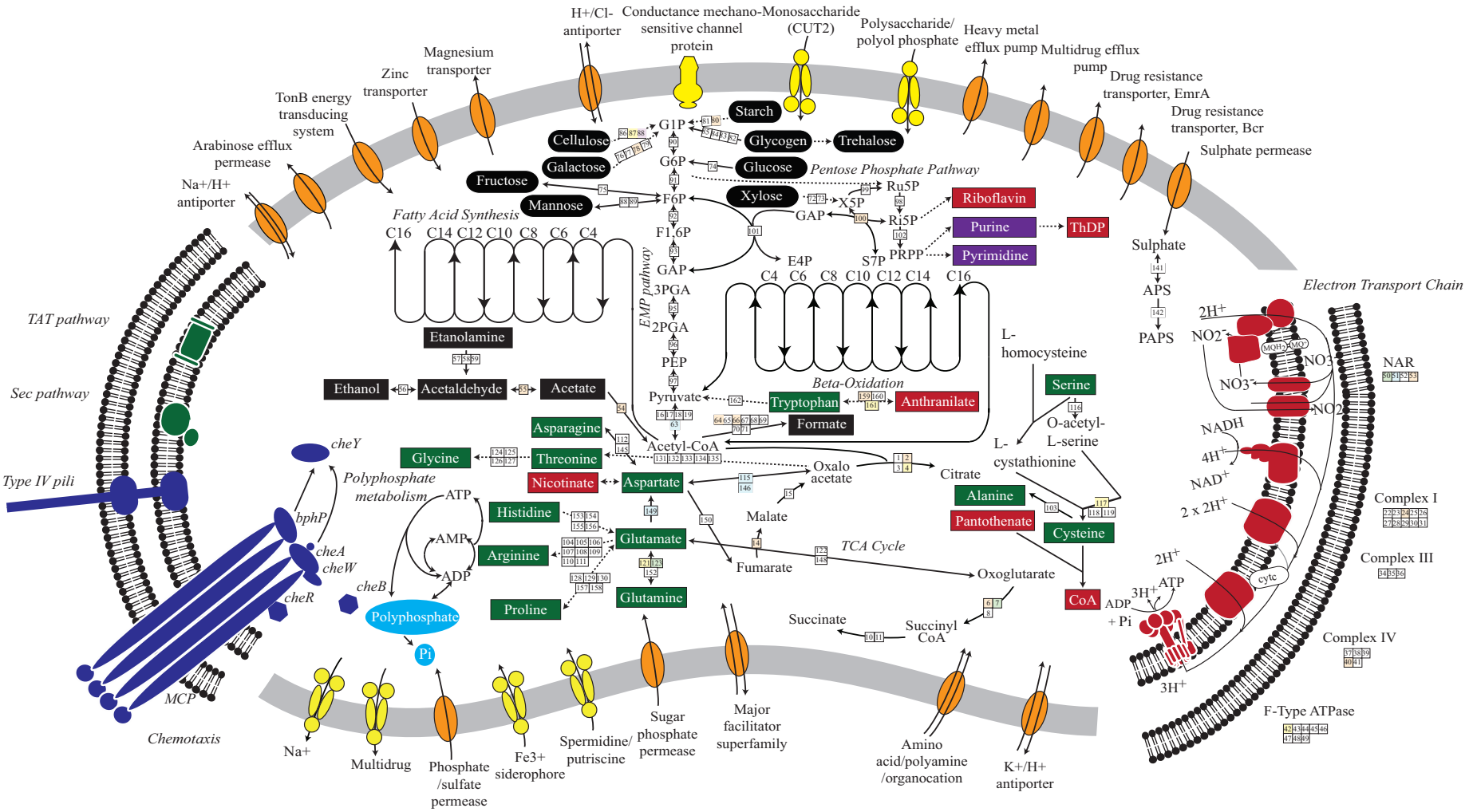


Figure 4.3. Metabolic reconstruction of P3DObs2



Metabolic predictions for P3DObs1 and 2 are based on genes annotated by IMG/ER [24]. Solid and dashed lines represent single or multiple steps in a pathway respectively. Black ovals indicate substrates that enter the glycolysis pathway. Fermentation end-products are indicated as black rectangles. Both genomes are capable of oxidative phosphorylation as they contain a complete TCA cycle and electron transport chain. Biosynthetic products are shown in green (amino acids), red (co-factors and vitamins), purple (nucleotides) and orange (non-mevalonate pathway products). ATP-binding cassette transporters are highlighted in yellow and permeases, pumps and transporters are highlighted in orange. The direction of substrate transport across the membrane is shown with arrows. P3DObs1 and 2 are missing all recognised photosynthesis genes including those for Photosystems I and II, chlorophyll and antennae proteins. Numbers in boxes refer to the enzymes (**Table S4.1**) required for the metabolic pathways in white boxes (FPKG 0), orange boxes (FPKG >0-20), yellow boxes (FPKG >20-50), green boxes (FPKG >50-100), blue boxes (FPKG >100-500), purple boxes (>500-1000) and red boxes (>1000). If multiple copies are identified, the box colours are based on the most highly expressed.

4.4.5.3 Amino acid metabolism

Both P3DObs genomes encode pathways for alanine, arginine, asparagine, aspartate, cysteine, glutamate, glutamine, glycine and proline synthesis (**Figures 4.2** and **4.3**). In addition, P3DObs2 is able to synthesise threonine. Both genomes are able to synthesise chorismate, which serves as a precursor of aromatic amino acids (phenylalanine, tyrosine and tryptophan). However, only the tryptophan biosynthesis pathway is complete for P3DObs1. Multiple amino acid transporters are present in both genomes that may complement amino acid biosynthesis in these bacteria.

The genes for the sulphate activation pathway (sulphate adenylyltransferase, adenylysulphate kinase, 2'-phospho-adenylylsulphate reductase and sulphite reductase), as well as sulphate permeases are encoded on Obs2 and half the genes required are encoded on Obs1, indicating that Obs2 has the potential to use sulphate [46]. The sulphate activation pathway is also found in the Oxyphotobacteria. In addition, both genomes contain serine O-acetyltransferase and cysteine synthase for biosynthesis of cysteine from serine and sulphide. The sulphate genes are not congruent with the Oxyphotobacteria, suggesting independent gain.

Both bacteria have the potential to catabolise asparagine to aspartate, aspartate to malate, glutamate to aspartate and fumarate, glutamine to glutamate, histidine to glutamate, proline to glutamate and tryptophan to pyruvate and anthranilate, which could be used as a source of nitrogen for the bacteria. In addition, P3DObs1 is able to catabolise serine and P3DObs2 is able to catabolise alanine.

4.4.5.4 Nucleotide, coenzyme and cofactor biosynthesis

Both genomes contain the genes required for the synthesis of purines and pyrimidines, as well as class II (coenzyme B_{12}-dependent) ribonucleoside-diphosphate reductase, which can covert ribonucleotides to deoxyribonucleotides. P3DObs1 and 2 are able to synthesise pantothenate and riboflavin, with P3DObs1 also able to synthesise biotin and co-enzyme A and P3DObs2 able to synthesise thiamine diphosphate.

4.4.5.5 Fatty acid biosynthesis and beta-oxidation

Both bacteria are capable of fatty acid biosynthesis and beta-oxidation with fatty acid desaturases, fatty acid hydroxylases and cyclopropane fatty-acyl-phospholipid synthases for the synthesis of branched-chain fatty acids. Fatty acids can be degraded to the two-carbon compound acetyl-CoA, which can then be re-routed to the TCA cycle for complete catabolism or for P3DObs1 routed into the glyoxylate shunt for biosynthesis.

4.4.6 Chemotaxis and motility

P3DObs1 and 2 contain multiple genes required for the assembly of type IV pili (*tadBDEG*, *cpaABF, pilBFMT)* but genes for flagella are absent. They also contain multiple genes for chemotaxis that may help them to sense environmental cues. Components of the signaling system for the *Obscuribacterales* include the two-component systems with histidine kinase and receiver domains and the cyclic diguanylate signaling system composed of enzymatic domains, GGDEF and EAL. Both genomes had the largest number of ligand-binding PAS domain S-boxes found in the Melainabacteria, with 51 in P3DObs1 and 61 in P3DObs2. The median number for the Melainabacteria is 5.5. PAS domains monitor changes in light, redox potential, oxygen, small ligands, and the overall energy level of a cell and help bacteria to adapt to changing environments [47]. In addition, signal transduction histidine kinase EnvZ-like, dimerisation/phosphoacceptor domains were found in many of the PAS domain S-boxes. These have been shown to play a central role in osmoregulation as sensor-transmitters in two-component systems with OmpR [48]. Multiple GAF domains were also identified, which can be used to bind cyclic nucleotides.

4.4.7 Antibiotics and secondary metabolites

Multiple bacteriocins were identified in the two palsa genomes, however none of the bacteriocins were similar to those found in the Bactibase database [30] indicating their novelty. Both genomes also contain a type III polyketide synthase (PKS) (a predicted nargingenin-chalcone synthase) and two acyl transferase domains in PKS enzymes (**Table S4.3**). This suggests that the Obscuribacterales are equipped to attack other microbial species in their surroundings.

4.4.8 Secretory systems

P3DObs2 contains genes required for the major secretory pathway, the general Secretion route (Sec-pathway). It has the *secYEG* genes*,* which are used to construct the protein conducting channel and *secA* gene, the peripheral associated ATPase which drives the translocation of the secretory proteins across the membrane. Both P3DObs1 and 2 contain *tatAC* genes required to form the receptor and protein conducting machinery for the Twin-arginine translocation pathway (Tat-pathway). The Tat-pathway allows for secretory proteins to be translocated across the cytoplasmic membrane in a folded conformation [49]. A total of 647 (9.7%) of the total protein coding genes in P3DObs1 encode signal peptides, whereas in P3DObs2, 682 (11.1%) of the total protein coding genes encode signal peptides.

4.4.9 Drug and antibiotics resistance

Both genomes contain genes for ABC-type multidrug transport systems, multidrug efflux pumps, drug resistance transporters and permeases of the drug/metabolite transporter superfamily. These transporters can be used to export drugs out of the genomes. In addition, both genomes contain a large number of beta-lactamases from classes A, C and D, which could provide resistance to a range of beta-lactam antibiotics [50].

4.4.10 Potential adaptations to a cold climate

4.4.10.1 Sigma factors

Three copies of σ^{70} factor, *rpoD*, the house-keeping and general stress response sigma factor, were identified in P3DObs1 and one was identified in P3DObs2. Multiple copies of *rpoD* have been described in other bacteria from cold climates such as *Planococcus halocryophilus* [51], *Psychromonas ingrahamii* [52] and *Arthrobacter* isolates from Antarctic soil [53]. In addition, a copy of σ^{24}, *rpoE*, associated with regulating cellular responses to heat shock and other stresses was identified in P3DObs1. Both *rpoD* and *rpoE* were expressed in the P3DObs genomes (**Table S4.1**). In comparison, *Ca.* O. phosphatis contains two copies of *rpoD* and no *rpoE*.

4.4.10.2 Chaperones and stress proteins

Surprisingly, cold shock proteins in both genomes were absent but multiple chaperone proteins are encoded on the genomes that may help with folding of proteins in cold conditions. Both genomes contained the chaperones ClpB, DnaJ, DnaK and GrpE, which together have shown to repress and reverse stress-damaged proteins from an aggregated state [54, 55]. In addition, P3DObs2 also has GroES and GroL. P3DObs1 and 2 contain a large number of copies of the nucleotide-binding universal stress protein, uspA (20 and 27 respectively), whereas *Ca.* O. phosphatis only has 10. The universal stress protein can be stimulated by a range of environmental conditions including cold shock, leading to repression of *uspA* and growth arrest [56]. One quarter of the *uspA* genes were expressed by P3DObs1 and 59% were expressed by P3DObs2. Furthermore, these genes were amongst the most highly expressed genes in both genomes.

4.4.10.3 Transcription and translation

DNA/RNA replication may be maintained at low temperatures by helicases belonging to the DEAD-box proteins, RNA-dependent ATPases that unwind short RNA duplexes [57] and recombination factors RecA/Q. These helicases and recombination factors have not been found in *Ca.* O. phosphatis. Transcription and translation factors that may help the palsa genomes include transcription termination (NusA/B), translation initiation factors (IF2, IF3) and elongation factor Ef-

tu. P3DObs1 did not show any expression for the transcription and translation factors but P3DObs2 had low expression for NusB and translation initiation factor 2.

4.4.10.4 Carbon and energy reserves

As found in *Ca.* O. phosphatis, P3DObs2 has the potential for polyphosphate metabolism, containing the polyphosphate kinase 1 and 2 (*ppk1* and *ppk2*), exopolyphosphatase (*ppx*) and polyphosphate:nucleotide phosphotransferase, adenylate kinase (*adk*) and phosphate/sulphate permeases [58]. The polyphosphate may be used as a carbon store which may provide a competitive advantage to these bacteria upon thawing of the palsa. P3DObs1 contains the genes *ppk1*, *ppk2*, polyphosphate:nucleotide phosphotransferase, PPK2 family and phosphate/sulphate permeases but is missing the essential exopolyphosphatase. Only the *ppk2* was expressed in both genomes.

4.4.10.5 Cryoprotectants

Trehalose protects cells from desiccation, osmotic stress and cold shock, as well as serve as a carbon source during nutrient starvation [59]. It is thought to have a colligative effect (due to particle number rather than nature) but may also help in preventing protein denaturation and aggregation [6]. Both Obscuribacterales genomes are capable of producing trehalose from maltodextrin. In addition, P3DObs2 contains an ABC-type transporter system that can transport proline/glycine betaine and choline and P3DObs1 contains a choline dehydrogenase, which can catalyse the chemical reaction of choline to betaine aldehyde. Glycine betaine and its precursor choloine is a reported osmoprotectant and has been found in other cold-stressed organisms, such as *Colwellia psychrerythraea* 34H [10].

4.4.10.6 Oxidative stress

At low temperatures, the solubility of gases increases and there is an accumulation of reactive oxygen species leading to a greater potential for oxidative damage than in temperate habitats [60]. The Obscuribacterales genomes contain several genes encoding a variety of enzymes to combat free radical damage including catalases, peroxiredoxin and superoxide dismutases for defense against reactive oxygen species (ROS). One peroxiredoxin, reduces hydrogen peroxide, was expressed at high levels in both genomes and one superoxide dismutase was expressed in P3DObs1.The genes for the ROS are not congruent with the Oxyphotobacteria suggesting that the oxidative phosophorylation pathway wasa independently gained in the Obscuribacterales.

4.4.10.7 Cell membrane adaptations

A decrease in temperature can lead to a reduction in membrane fluidity, a transition to gel-phase which can finally lead to a loss of function. In cold temperatures, organisms can produce a higher content of unsaturated, polyunsaturated and methyl branched fatty acids [61]. This adaptation is thought to lead to an increase in membrane fluidity. The Obscuribacterales genomes contain multiple copies of fatty acid desaturases, which suggests that fatty acids can be unsaturated.

4.5 Conclusion

Our findings expand the knowledge of the newly recognised class of non-photosynthetic Cyanobacteria, the Melainabacteria and in particular the class Obscruibacterales. Both genomes are psychrotrophic, Gram-negative, rod-shaped, facultative anaerobes and are slightly larger than *Ca*. O. phosphatis. Both are capable of using fermentation and respiration, as found in *Ca.* O. phosphatis.

The identification of cellulases produced by the cold-adapted Obscuribacterales genomes from P3D could potentially be used for environmental bioremediation, the food industry and molecular biology due to its predicted stability at low temperatures [62].

4.6 Acknowledgments

We thank Carmody McCalley for collecting P3D and the IsoGenie team for funding. We thank Paul Evans and Serene Lowe for extracting the samples and Serene Lowe for preparing the DNA for sequencing at IMB and ACE, UQ.

4.7 References

1. van Everdingen, R., Multi-Language Glossary of Permafrost and Related Ground-Ice Terms, Version 2.0., ed. R.O. van Everdingen. 1998, University of Calgary, Calgary, Canada.
2. Zhang, T., et al. Distribution of seasonally and perennially frozen ground in the Northern Hemisphere. in Proceedings of the 8th International Conference on Permafrost. 2003. AA Balkema Publishers.
3. Steven, B., et al., Microbial ecology and biodiversity in permafrost. Extremophiles, 2006. **10**(4): p. 259-67.

4. Qiu, Y., S. Kathariou, and D.M. Lubman, Proteomic analysis of cold adaptation in a Siberian permafrost bacterium – *Exiguobacterium sibiricum* 255–15 by two-dimensional liquid separation coupled with mass spectrometry. PROTEOMICS, 2006. **6**(19): p. 5221-5233.
5. Chong, B.E., et al., Use of non-porous reversed-phase high-performance liquid chromatography for protein profiling and isolation of proteins induced by temperature variations for Siberian permafrost bacteria with identification by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and capillary electrophoresis–electrospray ionization mass spectrometry. Journal of Chromatography B: Biomedical Sciences and Applications, 2000. **748**(1): p. 167-177.
6. Phadtare, S., Recent developments in bacterial cold-shock response. Current Issues in Molecular Biology, 2004. **6**(2): p. 125-36.
7. Méndez, M.B., et al., Novel Roles of the Master Transcription Factors Spo0A and σB for Survival and Sporulation of Bacillus subtilis at Low Growth Temperature. Journal of Bacteriology, 2004. **186**(4): p. 989-1000.
8. Shcherbakova, V.A., et al., Novel psychrophilic anaerobic spore-forming bacterium from the overcooled water brine in permafrost: description *Clostridium algoriphilum* sp. nov. Extremophiles, 2005. **9**(3): p. 239-246.
9. Bayles, D.O. and B.J. Wilkinson, Osmoprotectants and cryoprotectants for *Listeria monocytogenes*. Letters in Applied Microbiology, 2000. **30**(1): p. 23-27.
10. Methe, B.A., et al., The psychrophilic lifestyle as revealed by the genome sequence of *Colwellia psychrerythraea* 34H through genomic and proteomic analyses. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(31): p. 10913-8.
11. Steven, B., et al., Characterization of the microbial diversity in a permafrost sample from the Canadian high Arctic using culture-dependent and culture-independent methods. FEMS Microbiology Ecology, 2007. **59**(2): p. 513-23.
12. Jansson, J.K. and N. Tas, The microbial ecology of permafrost. Nature Reviews Microbiology, 2014. **12**(6): p. 414-425.
13. Vishnivetskaya, T., Viable Cyanobacteria and Green Algae from the Permafrost Darkness, in Permafrost Soils, R. Margesin, Editor. 2009, Springer Berlin Heidelberg. p. 73-84.
14. Blanco, Y., et al., Prokaryotic communities and operating metabolisms in the surface and the permafrost of Deception Island (Antarctica). Environmental microbiology, 2012. **14**(9): p. 2495-2510.

15. Di Rienzi, S.C., et al., The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. eLife, 2013. **2**: p. e01102. Di
16. Soo, R.M., et al., An Expanded Genomic Representation of the Phylum Cyanobacteria. Genome Biology and Evolution, 2014. **6**(5): p. 1031-1045.
17. Soo, R.M., et al., Back from the dead; the curious tale of the predatory cyanobacterium Vampirovibrio chlorellavorus. PeerJ, 2015.
18. McCalley, C.K., et al., Methane dynamics regulated by microbial community response to permafrost thaw. Nature, 2014. **514**(7523): p. 478-481.
19. McDonald, D., et al., An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J, 2012. **6**(3): p. 610-8.
20. Markowitz, V.M., et al., IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Research, 2014. **42**: p. D568-73.
21. Parks, D.H., et al., CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. PeerJ PrePrints, 2014. **2**: p. e554v1.
22. Finn, R.D., J. Clements, and S.R. Eddy, HMMER web server: interactive sequence similarity searching. Nucleic Acids Research, 2011. **39**: p. W29-37.
23. Price, M.N., P.S. Dehal, and A.P. Arkin, FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE, 2010. **5**(3): p. e9490.
24. Markowitz, V.M., et al., IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics, 2009. **25**(17): p. 2271-2278.
25. Seemann, T., Prokka: rapid prokaryotic genome annotation. Bioinformatics, 2014.
26. Hüttenhofer, A., P. Schattner, and N. Polacek, Non-coding RNAs: hope or hype? Trends in Genetics, 2005. **21**(5): p. 289-297.
27. Yin, Y., et al., dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Research, 2012. **40**: p. W445-51.
28. Grissa, I., G. Vergnaud, and C. Pourcel, CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Research, 2007. **35**: p. W52-7.
29. Medema, M.H., et al., antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Research, 2011.
30. Hammami, R., et al., BACTIBASE second release: a database and tool platform for bacteriocin characterization. BMC Microbiology, 2010. **10**(1): p. 22.

31. Caspi, R., et al., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Research, 2014. **42**(D1): p. D459-D471.
32. Kopylova, E., L. Noe, and H. Touzet, SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics, 2012. **28**(24): p. 3211-7.
33. Quast, C., et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 2013. **41**(D1): p. D590-D596.
34. Milne, I., et al., Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics, 2013. **14**(2): p. 193-202.
35. Médigue, C., et al., Coping with cold: The genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. Genome Research, 2005. **15**(10): p. 1325-1335.
36. Luo, C., L.M. Rodriguez-R, and K.T. Konstantinidis, MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. Nucleic Acids Research, 2014: p. gku169.
37. Hoiczyk, E. and A. Hansel, Cyanobacterial Cell Walls: News from an Unusual Prokaryotic Envelope. Journal of Bacteriology, 2000. **182**(5): p. 1191-1199.
38. White, C.L., A. Kitich, and J.W. Gober, Positioning cell wall synthetic complexes by the bacterial morphogenetic proteins MreB and MreD. Molecular Microbiology, 2010. **76**(3): p. 616-633.
39. Alber, B.E., et al., Study of an alternate glyoxylate cycle for acetate assimilation by *Rhodobacter sphaeroides*. Molecular Microbiology, 2006. **61**(2): p. 297-309.
40. Kulajta, C., et al., Multi-step assembly pathway of the *cbb*$_3$-type cytochrome c oxidase complex. Journal of Molecular Biology, 2006. **355**(5): p. 989-1004.
41. Garsin, D.A., Ethanolamine Utilization in Bacterial Pathogens: Roles and Regulation. Nature reviews Microbiology, 2010. **8**(4): p. 290-295.
42. Vignais, P.M. and B. Billoud, Occurrence, classification, and biological function of hydrogenases: an overview. Chemical reviews, 2007. **107**(10): p. 4206-4272.
43. Schmitz, O., et al., HoxE—a subunit specific for the pentameric bidirectional hydrogenase complex (HoxEFUYH) of cyanobacteria. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 2002. **1554**(1–2): p. 66-74.
44. Tamagnini, P., et al., Hydrogenases and Hydrogen Metabolism of Cyanobacteria. Microbiology and Molecular Biology Reviews, 2002. **66**(1): p. 1-20.
45. Kruger, N.J. and A. von Schaewen, The oxidative pentose phosphate pathway: structure and organisation. Current Opinion in Plant Biology, 2003. **6**(3): p. 236-246.

46. Leyh, T.S., J.C. Taylor, and G.D. Markham, The sulfate activation locus of *Escherichia coli* K12: cloning, genetic, and enzymatic characterization. Journal of Biological Chemistry, 1988. **263**(5): p. 2409-2416.
47. Taylor, B.L. and I.B. Zhulin, PAS domains: internal sensors of oxygen, redox potential, and light. Microbiology and Molecular Biology Reviews, 1999. **63**(2): p. 479-506.
48. Forst, S., J. Delgado, and M. Inouye, Phosphorylation of OmpR by the osmosensor EnvZ modulates expression of the *ompF* and *ompC* genes in *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(16): p. 6052-6056.
49. Natale, P., T. Brüser, and A.J.M. Driessen, Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—Distinct translocases and mechanisms. Biochimica et Biophysica Acta (BBA) - Biomembranes, 2008. **1778**(9): p. 1735-1756.
50. Bush, K., G.A. Jacoby, and A.A. Medeiros, A functional classification scheme for beta-lactamases and its correlation with molecular structure. Antimicrobial Agents and Chemotherapy, 1995. **39**(6): p. 1211-1233.
51. Mykytczuk, N.C., et al., Bacterial growth at -15 degrees C; molecular insights from the permafrost bacterium *Planococcus halocryophilus* Or1. ISME J, 2013. **7**(6): p. 1211-26.
52. Riley, M., et al., Genomics of an extreme psychrophile, *Psychromonas ingrahamii*. BMC Genomics, 2008. **9**: p. 210.
53. Dsouza, M., et al., Genomic and phenotypic insights into the ecology of *Arthrobacter* from Antarctic soils. BMC Genomics, 2015. **16**: p. 36.
54. Zolkiewski, M., ClpB cooperates with DnaK, DnaJ, and GrpE in suppressing protein aggregation. A novel multi-chaperone system from *Escherichia coli*. Journal of Biological Chemistry, 1999. **274**(40): p. 28083-6.
55. Lee, S., et al., The structure of ClpB: a molecular chaperone that rescues proteins from an aggregated state. Cell, 2003. **115**(2): p. 229-40.
56. Kvint, K., et al., The bacterial universal stress protein: function and regulation. Current Opinion in Microbiology, 2003. **6**(2): p. 140-5.
57. Cartier, G., et al., Cold Adaptation in DEAD-Box Proteins. Biochemistry, 2010. **49**(12): p. 2636-2646.
58. Seviour, R.J. and P.H. Nielsen, Microbial Ecology of Activated Sludge. 2010: IWA Publishing.
59. Kandror, O., A. DeLeon, and A.L. Goldberg, Trehalose synthesis is induced upon exposure of *Escherichia coli* to cold and is essential for viability at low temperatures. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(15): p. 9727-32.

60. Chattopadhyay, M.K., Mechanism of bacterial adaptation to low temperature. Journal of Biosciences, 2006. **31**(1): p. 157-165.
61. D'Amico, S., et al., Psychrophilic microorganisms: challenges for life. EMBO Reports, 2006. **7**(4): p. 385-389.
62. Kasana, R.C. and A. Gulati, Cellulases from psychrophilic microorganisms: a review. Journal of Basic Microbiology, 2011. **51**(6): p. 572-579.

Chapter 5: Conclusion and future directions

5.1 Overview

Uncultured basal cyanobacteria have been identified in multiple 16S rRNA-based culture-independent studies originating from a range of environments including mammal guts [1-6], anaerobic digestors [7], marine sediments [8] and drinking water [9]. These 16S rRNA sequences were originally classified as a cyanobacterial class in the Greengenes [10] and Silva [11] databases. Many of the environments where the sequences were obtained are aphotic, leading to the question of whether these basal Cyanobacteria are non-photosynthetic and if so, should they be classified as Cyanobacteria or as a separate phylum. With the advent of metagenomics and improved binning methods (Chapter 1), Di Rienzi and colleagues were able to obtain six basal cyanobacteria population genomes from metagenomic samples collected from human gut and a groundwater aquifer. Most notably, these genomes lack the photosynthetic apparatus needed for oxygenic photosynthesis, instead apparently obtaining their energy via fermentative pathways. Based on this conspicuous physiological difference from photosynthetic cyanobacteria, Di Rienzi *et al.* (2013) proposed that the lineage should be classified as a sister phylum, the Melainabacteria (dark nymph), rather than a class within the Cyanobacteria.

The main aim of my thesis was to characterise members of the Melainabacteria through genome sequencing and transcriptomics, and to assess whether they should be classified within the phylum Cyanobacteria or be defined as a sister phylum. In Chapter 2, population genomes were obtained from koala faeces, human faeces, an enhanced biological phosphorous removal (EBPR) bioreactor and an upflow sludge blanket reactor which expanded the number of near-complete Melainabacteria genomes from six [12] to eleven [13]. Phylogenetic analyses, as well as genes that are found only in Cyanobacteria, suggest that the Melainabacteria should be classified within the Cyanobacteria. This is based on the Melainabacteria being robustly monophyletic and sharing a number of uniquely cyanobacterial traits with the photosynthetic Cyanobacteria, including unusual cell envelope components for Gram-negative bacteria, putative circadian rhythm and light response regulators. Additionally, we classified all photosynthetic cyanobacteria in a single class called the Oxyphotobacteria, a name originally proposed by Gibbons and Murray (1978) for photosynthetic cyanobacteria. Di Rienzi *et al.* (2013) initially concluded that the Melainabacteria are all fermentative, but our broader sampling of this lineage, shows that at least one group, the Obscuribacterales, is capable of respiration. During analysis of Melainabacteria 16S rRNA

sequence data, a putative cultured representative was discovered in the Greengenes database, a predatory bacterium named *Vampirovibiro chlorellavorus*, misclassified as a Deltaproteobacterium in NCBI. In Chapter 3, genomic DNA was extracted from a 36-year old vial of lyophilised *V. chlorellavorus* co-cultured with its microalgal prey, and shotgun-sequenced. A high quality draft genome of *V. chlorellavorus* was assembled from these data and phylogenetic analyses of multiple concatenated marker genes confirmed its affiliation with the Cyanobacteria. A detailed model of its predatory lifestyle was inferred from a metabolic reconstruction revealing that this organism likely uses a type IV secretion system to attach to its prey. In Chapter 4, a permafrost sample containing a member of the order *Obscuribacterales* was identified and two population genomes were extracted from a single permafrost metagenome. Metabolic reconstruction of the *Obscuribacterales* genomes identified multiple genes that would provide the genomes with the ability to survive in cold environments. In addition, metatranscriptomics was used to identify genes that were expressed by the genomes in the permafrost sample.

5.2 Definition of a phylum

Traditionally taxonomists have focused on defining microorganisms at the species-level but little emphasis has been placed at the phylum-level. The question of how to define a phylum arose while writing Chapter 2 as Di Rienzi had classified the Melainabacteria as a new sister phylum, however we believe that the Melainabacteria are a basal lineage of the Cyanobacteria. The most widely used gene for identifying phylogenetic placement, including phylum, is the 16S rRNA gene, which is sufficiently conserved and ubiquitous enough to reveal ancient relationships [15]. Di Rienzi and colleagues suggested that the Melainabacteria are a new sister phylum to the Cyanobacteria based on the Melainabacteria sharing no more than 84% identity to the cyanobacterial 16S rRNA (the recommended threshold is 85% based on the definition by Hugenholtz *et al.,* 1998). However, the main criterium that the lineage is reproducibly monophyletic and unaffiliated with all other division-level relatedness groups, were not met. More importantly, this definition was established to deter single organisms being called a new phylum, which could occur with chimeric sequences (sequences composed of DNA from two or more parents) [10]. This case study highlights the somewhat arbitrary nature of defining phyla and a more systematic approach is required to make the definitions objective. Using hard thresholds is not ideal and evolutionary divergence should be taken into consideration.

With the advent of whole genome sequencing, it is now possible to define phyla based on more than one marker gene. For example, Average Nucleotide Identity (ANI) and Amino Acid Identity (AAI) can be used to measure the genetic relatedness between a pair of genomes by two-way BLAST [17,

18]. The values for AAI cut-offs are still arbitrary, as there is a grey area between different taxonomic ranks. Luo and colleagues defined an AAI cut-off of <45% for phylum, although the cut-offs range from 38-52% and the AAI cut-off for order is between 45-58% (7% overlap). However, they note that the 45% AAI cut-off encompasses deep-branching organisms that may be assigned to deep branching classes or domains as opposed to phyla in the future [19].

Using multiple conserved single copy marker genes to produce concatenated gene trees produces greater resolution than single gene trees as there is no reliance of an individual marker, allowing for differences in horizontal gene transfer [20, 21]. As the tree of life expands due to the decrease in cost and the efficiency of next-generation sequencers, defining phyla should improve but assessment may need to be taken on a phylum-to-phylum basis.

5.3 The evolution of photosynthesis in Cyanobacteria

Over 100 genes are devoted to the synthesis and regulation of the photosynthetic apparatus. Photosynthetic genes are found in small clusters (2 to 4 genes) throughout Oxyphotobacteria genomes [22]. Therefore if the Melainabacteria had been photosynthetic at one point in time, we may expect to see remnants of these photosynthetic gene clusters. However, photosynthetic genes are completely absent from all Melainabacteria investigated to date, suggesting that they diverged from the Oxyphotobacteria prior to acquisition of any photosynthetic apparatus, or possibly as a result of the acquisition in the oxyphotobacterial ancestor [12].

Recent studies have suggested the opposite, that the Melainabacteria/Oxyphotobacteria ancestor was photosynthetic. Cardona (2014) reconstructed phylogenetic trees for type I (RC1) (Cyanobacteria, Heliobacteria, Acidobacteria and Chlorobi) and type II reaction centre (RC2) (Cyanobacteria, Proteobacteria and Chloroflexi) subunits, as well as 16S rRNA gene trees. From the RC1 and RC2 phylogenetic trees he suggested that all RC1 subunits originated from one ancestral protein that underwent several episodes of gene duplication. This is also the case for the RC2 subunits. However, the two different proteins are homologous and originated from a more ancient reaction centre subunit. He concludes that the phylogenetic trees show that the divergence of the ancestral RC1 and RC2 proteins occurred before the diversification of the extant groups of phototrophic bacteria. For example, Acidobacteria which contains RC1 proteins, clusters within the Proteobacteria that contain RC2 proteins. Furthermore, the evolution of proteins involved in the synthesis of chlorophyll also supports this relationship. He posits that reaction centres and pigments evolved early in the evolution of bacteria, if not before their last common ancestor, and then diversified as the numerous bacteria phyla expanded and questions whether the invention of

photosynthesis triggered the explosion of bacterial diversity during the Archean age. In addition, he posits that even though water oxidation evolved after the Melainabacteria/Oxyphotobacteria divergence, it is more likely that the Melainabacteria lost their reaction centres [23]. Harel and colleagues (2015) constructed protein similarity networks from four metabolic groups (aerobes, anaerobes, phototrophs, and methanogens) and the Cyanobacteria (but did not include the Melainabacteria) to understand the evolutionary history and gene content of primordial Cyanobacteria. They found that the Cyanobacteria share the largest number of unique genes with obligate anaerobes followed by obligate aerobes suggesting either a retention of anaerobic ancestry in Cyanobacteria or a high amount of horizontal gene transfer between Cyanobacteria and obligate anaerobes relative to other bacteria. In addition, they identified 15 core cyanobacterial functions in three genomes of the Melainabacteria, with fourteen of these functions detected in other photosynthetic groups. From these inferences, Harel and colleagues raised the possibility that the Melainabacteria and Cyanobacteria shared a photosynthetic common ancestor. However, the 15 core functions identified in the Melainabacteria were not used for photosynthesis and no non-photosynthetic outgroups were used to determine whether these genes are part of a bacterial core, raising doubts about this conclusion. Similarly, for the Cardona hypothesis to be correct, all non-photosynthetic bacteria would have had to lose their ancestral photosynthetic genes which is less parsimonious than individual lineages gaining photosynthetic genes through a combination of vertical descent and lateral gene transfer.

Due to the antiquity of the event, it may not be possible to determine whether the Melainabacteria ancestor was photosynthetic or not. If the ancestor was indeed photosynthetic and the Melainabacteria diverged from the Oxyphotobacteria as a result of loss of photosynthesis due to colonisation of aphotic habitats, enough time has elapsed that all remnants of ancestral photosynthetic genes would have been purged from modern Melainabacteria. Sequencing of representatives from the third class of Cyanobacteria, ML635J-21, may serve to shed light on the nature of the cyanobacterial ancestor.

5.4 The evolution of respiration in Melainabacteria

Di Rienzi *et al.* (2014) suggested that the Melainabacteria may have remained restricted to anoxic habitats since diverging from the Oxyphotobacteria, however, we identified Melainabacteria with genes for respiration occupying oxic or periodically oxic habitats. *Vampirovibrio chlorellavorus* (Chapter 3) has the potential to respire aerobically in freshwater whereas members of the order *Obscuribacterales* (Chapters 2 and 4) contain genes for both aerobic and anaerobic respiration. In

addition, many of the Melainabacteria clones found in the Greengenes and Silva databases have been isolated from aerobic environments, including soil [25-27] and drinking water [9].

Most members of the order *Gastranaerophilales* have been identified in mammalian guts. However, it has been estimated that mammals arose during the Middle Jurassic period, 176 Mya [28], indicating that mammal guts have been subsequently colonised by the *Gastranaerophilales*. Therefore, it could be suggested that the Melainabacteria may have began as fermenters in an anaerobic environment and then some may have gained the genes required for life in oxic habitats as the planet became aerobic from the photosynthetic Oxyphotobacteria, whereas others remained living in anoxic environments.

5.5 Future directions

Future directions for investigating the class Melainabacteria include designing primers to target the 16S rRNA genes of the Melainabacteria, which could lead to an increased number of Melainabacteria representatives being discovered in different ecological niches. Multiple primers (PCR primers and real-time PCR primers) to target the Melainabacteria 16S rRNA genes at both the class- and order-level were designed and tested, however, these attempts were unsuccessful as the primers were non-specific (**Appendix D**). Perhaps designing primers to target regions of other conservative genes, for example housekeeping genes, or multiple genes, may lead to success.

Di Rienzi and colleagues suggested that members of the order *Gastranaerophilales* could be symbiotic due to the presence of hydrogenases. Fluorescence in situ hypbridisation (FISH) probes may help to identify if this order is symbiotic with methanogens. Multiple FISH probes were designed to target both the class Melainabacteria and three different orders (*Gastranaerophilales, Obscuribacterales* and *Vampirovibrionales*). FISH probes that target multiple regions of the 16S rRNA genes simultaneously, with the assistance of helper probes were used. However, like the 16S rRNA gene primers, they were found to be non-specific (**Appendix D**). Additionally, problems with autofluorescence from the environmental samples and the cells made it difficult to assess whether the FISH probes were working. Treatments to remove autofluorescence (H_2O_2 and toluidine blue O) were unsuccessful (**Appendix D**). As above, designing probes to target regions of other conservative genes or multiple genes may lead to success.

If the Melainabacteria could be cultured, perturbation studies combined with metatranscriptomics may provide a better understanding of the metabolism for this group and how they interact with other organisms and the environment. Information from the metabolic reconstructions created from

the metagenomics data should help to identify culturing media for the successful cultivation of Melainabacteria.

Further exploration into the function of the cytochromes in the Melainabacteria, may provide us with more information on how the Cyanobacteria have evolved from the ancestor of the Melainabacteria and the Cyanobacteria to be able to perform oxygenic photosynthesis. The phylogenetic tree from Chapter 2 that is included in **Figure S1.4**, suggests that another class (ML635J-21), more basal than the Melainabacteria may exist. Whether photosystems are present or absent in this lineage may shed more light on the debate of whether the Melainabacteria have gained or lost the genes for the photosystems.

5.6 References

1. Ley, R.E., et al., Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(31): p. 11070-11075.
2. Ley, R.E., et al., Microbial ecology: Human gut microbes associated with obesity. Nature, 2006. **444**(7122): p. 1022-1023.
3. Stecher, B., et al., *Salmonella enterica* serovar typhimurium exploits inflammation to compete with the intestinal microbiota. PLoS Biology, 2007. **5**(10): p. 2177-89.
4. Tajima, K., et al., Influence of high temperature and humidity on rumen bacterial diversity in Holstein heifers. Anaerobe, 2007. **13**(2): p. 57-64.
5. Monteils, V., et al., Potential core species and satellite species in the bacterial community within the rabbit caecum. FEMS Microbiology Ecology, 2008. **66**(3): p. 620-9.
6. Yang, S., et al., Bacterial diversity in the rumen of Gayals (*Bos frontalis*), Swamp buffaloes (*Bubalus bubalis*) and Holstein cow as revealed by cloned 16S rRNA gene sequences. Molecular Biology Reports, 2010. **37**(4): p. 2063-73.
7. Riviere, D., et al., Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. ISME J, 2009. **3**(6): p. 700-714.
8. Reed, D.W., et al., Microbial Communities from Methane Hydrate-Bearing Deep Marine Sediments in a Forearc Basin. Applied and Environmental Microbiology, 2002. **68**(8): p. 3759-3770.
9. Williams, M.M., et al., Phylogenetic diversity of drinking water bacteria in a distribution system simulator. Journal of Applied Microbiology, 2004. **96**(5): p. 954-964.
10. McDonald, D., et al., An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J, 2012. **6**(3): p. 610-8.

11. Quast, C., et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 2013. **41**(D1): p. D590-D596.
12. Di Rienzi, S.C., et al., The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. eLife, 2013. **2**: p. e01102. Di
13. Soo, R.M., et al., An Expanded Genomic Representation of the Phylum Cyanobacteria. Genome Biology and Evolution, 2014. **6**(5): p. 1031-1045.
14. Gibbons, N.E. and R.G.E. Murray, Validation of *Cyanobacteriales* Stanier in Gibbons and Murray 1978 as a New Order of the Kingdom *Procaryotae* Murray 1968, and of the Use of Neuter Plural Endings for *Photobacteria* and *Scotobacteria* classes nov. Gibbons and Murray 1978: Request for an Opinion. International Journal of Systematic Bacteriology, 1978. **28**(2): p. 332-333.
15. Woese, C.R. and G.E. Fox, Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America, 1977. **74**(11): p. 5088-5090.
16. Hugenholtz, P., B.M. Goebel, and N.R. Pace, Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology, 1998. **180**(18): p. 4765-74.
17. Konstantinidis, K.T. and J.M. Tiedje, Towards a genome-based taxonomy for prokaryotes. Journal of Bacteriology, 2005. **187**(18): p. 6258-6264.
18. Konstantinidis, K.T. and J.M. Tiedje, Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(7): p. 2567-2572.
19. Luo, C., L.M. Rodriguez-R, and K.T. Konstantinidis, MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. Nucleic Acids Research, 2014: p. gku169.
20. Ciccarelli, F.D., et al., Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science, 2006. **311**(5765): p. 1283-1287.
21. Wu, D., et al., A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature, 2009. **462**(7276): p. 1056-1060.
22. Shi, T., et al., Protein Interactions Limit the Rate of Evolution of Photosynthetic Genes in Cyanobacteria. Molecular Biology and Evolution, 2005. **22**(11): p. 2179-2189.
23. Cardona, T., A fresh look at the evolution and diversification of photochemical reaction centers. Photosynthesis Research, 2014: p. 1-24.
24. Harel, A., et al., Deciphering Primordial Cyanobacterial Genome Functions from Protein Network Analysis. Current Biology, 2015. **25**(5): p.628-634.

25. Elshahed, M.S., et al., Novelty and Uniqueness Patterns of Rare Members of the Soil Biosphere. Applied and Environmental Microbiology, 2008. **74**(17): p. 5422-5428.
26. Lesaulnier, C., et al., Elevated atmospheric CO_2 affects soil microbial diversity associated with trembling aspen. Environmental Microbiology, 2008. **10**(4): p. 926-941.
27. Cruz-Martinez, K., et al., Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. ISME J, 2009. **3**(6): p. 738-744.
28. Luo, Z.-X., Transformation and diversification in early mammal evolution. Nature, 2007. **450**(7172): p. 1011-1019.

Appendix A: Supplementary figures and tables for Chapter 2

Supplementary Figures

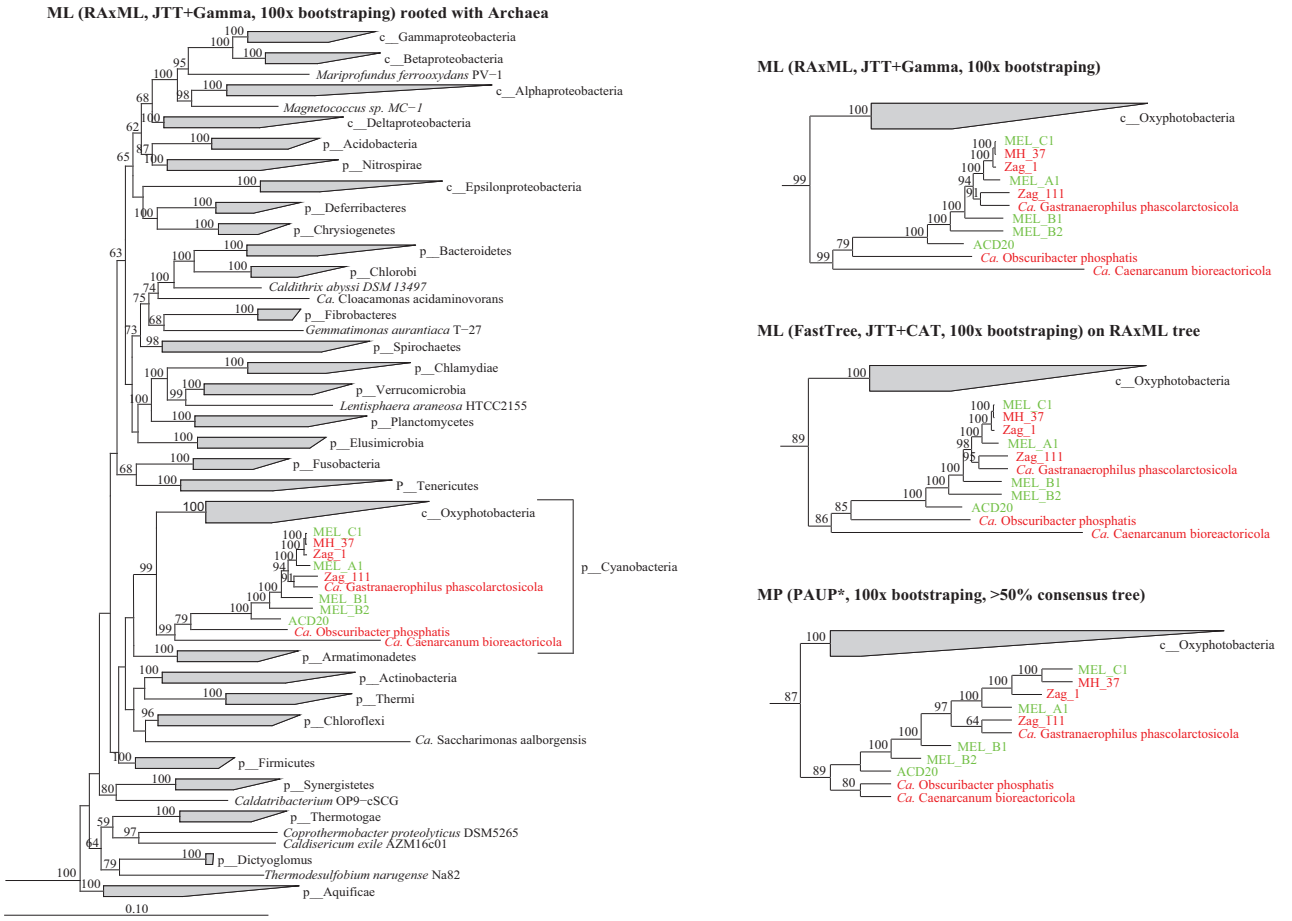


Figure S2.1. Phylogeny of Oxyphotobacteria and Melainabacteria among the bacterial phyla based on up to 38 marker genes

Phylogenetic maximum likelihood (RAxML, JTT, G) and maximum parsimony (PAUP*) trees based on a concatenated alignment of up to 38 marker genes, showing the phylogenetic robustness of the phylum Cyanobacteria and the classes Oxyphotobacteria and Melainabacteria. 422 OTUs (operational taxonomic units) from Bacteria and Archaea were used to produce the phylogenetic tree (**Table S2.4**). Bootstrap analyses (100 times) were performed for the data set with maximum likelihood (RAxML, JTT, G; FastTree, JTT, CAT) and maximum parsimony (PAUP*) methods, and the values obtained are shown in respective trees, except for the values from FastTree, which are shown on the tree generated with RAxML bootstrapping. Genomes in green are representatives from Di Rienzi *et al.,* 2013 and genomes in red are Melainabacteria from this study. Bootstrap values >60% are shown. p_ represents phylum and c_ represents class. *Candidatus* has been abbreviated to *Ca.* for the most complete genomes.

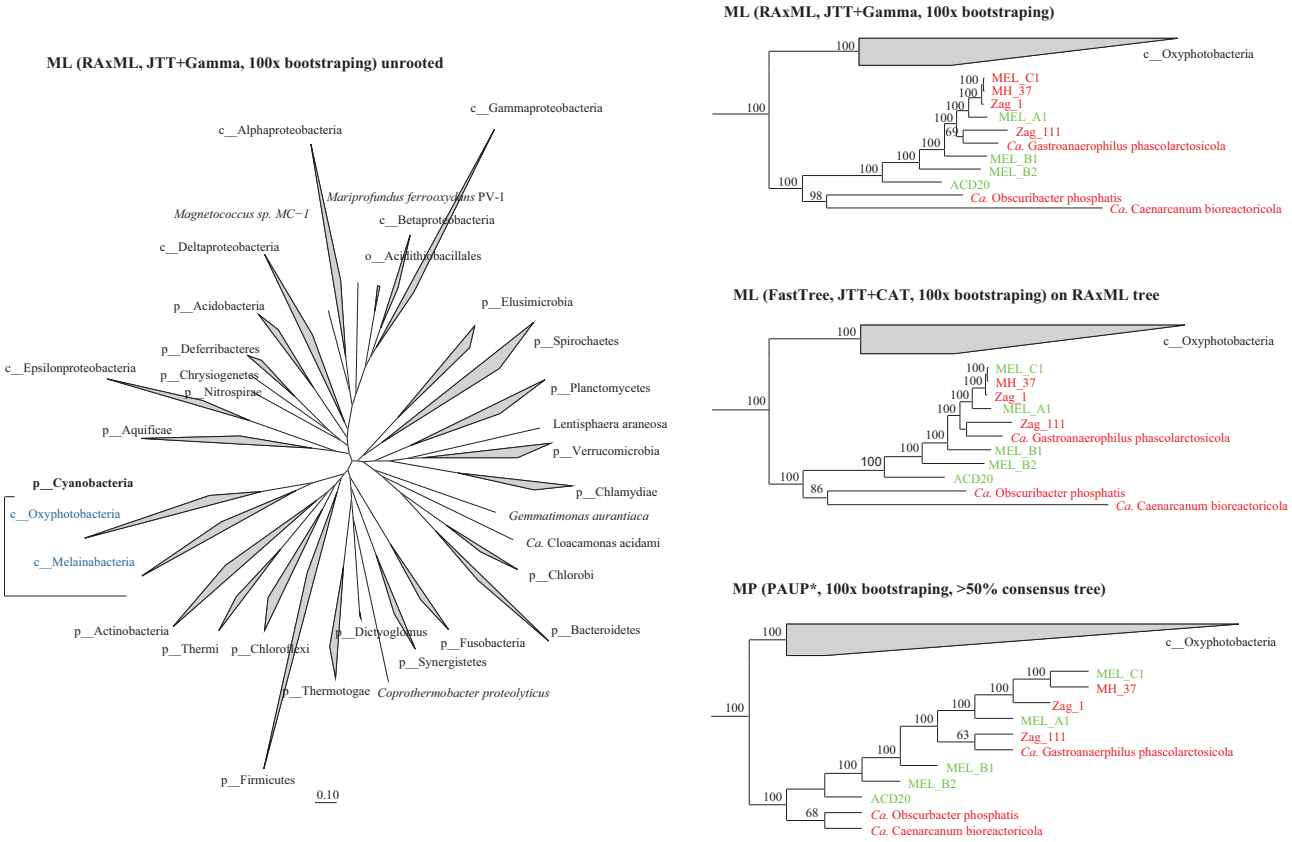


Figure S2.2. Phylogeny of Oxyphotobacteria and Melainabacteria among the bacterial phyla based on up to 83 marker genes

Phylogenetic maximum likelihood (RAxML, JTT, G) and maximum parsimony (PAUP*) trees based on a concatenated alignment of up to 83 marker genes, showing the phylogenetic robustness of the phylum Cyanobacteria and the classes Oxyphotobacteria and Melainabacteria. 322 OTUs from Bacteria were used to produce the phylogenetic tree (**Table S2.4**). Bootstrap analyses (100 times) were performed for the data set with maximum likelihood (RAxML, JTT, G; FastTree, JTT, CAT) and maximum parsimony (PAUP*) methods, and the values obtained are shown in respective trees, except for the values from FastTree, which are shown on the tree generated with RAxML. Genomes in green are representatives from Di Rienzi *et al.,* 2013 and genomes in red are Melainabacteria from this study. Bootstrap values >60% are shown. p_ represents phylum, c_ represents class and o_ represents order. *Candidatus* has been abbreviated to *Ca.* for the most complete genomes.

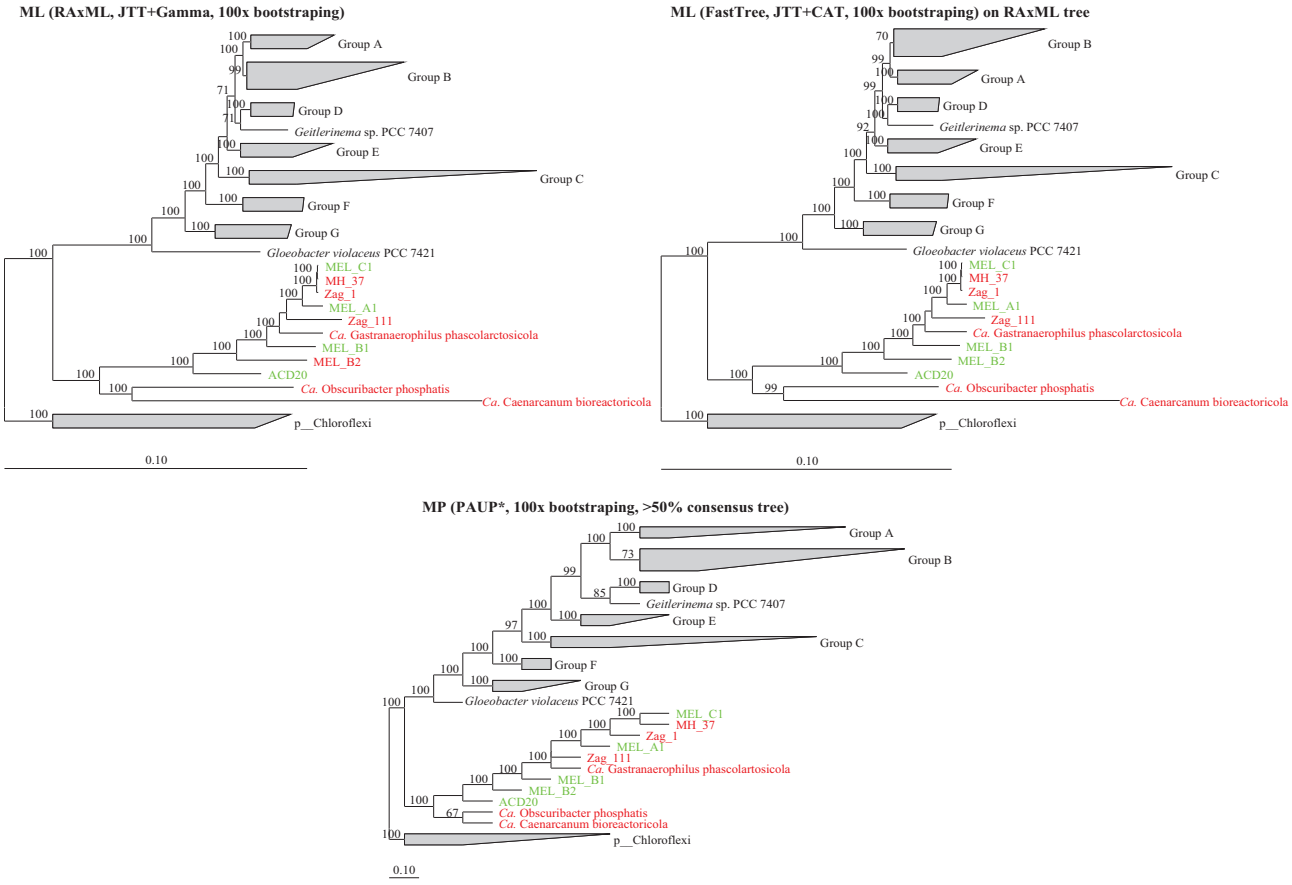


Figure S2.3. Phylogeny of Oxyphotobacteria and Melainabacteria in Cyanobacteria phylum based on up to 83 marker genes

Phylogenetic maximum likelihood (RAxML, JTT, G) and maximum parsimony (PAUP*) trees based on a concatenated alignment of up to 83 marker genes, showing the phylogenetic robustness of the phylum Cyanobacteria and the classes. The Oxyphotobacteria were grouped as according to Shih *et al.,* (2013) (**Table S2.3**). Bootstrap analyses (100 times) were performed for the data set with maximum likelihood (RAxML, JTT, G; FastTree, JTT, CAT) and maximum parsimony (PAUP*) methods, and the values obtained are shown in respective trees, except for the values from FastTree, which are shown on the tree generated with RAxML. Genomes in green are representatives from Di Rienzi *et al.,* 2013 and genomes in red are Melainabacteria from this study. Bootstrap values >60% are shown. p_ represents phylum, c_ represents class and o_ represents order. *Candidatus* has been abbreviated to *Ca.* for the most complete genomes.

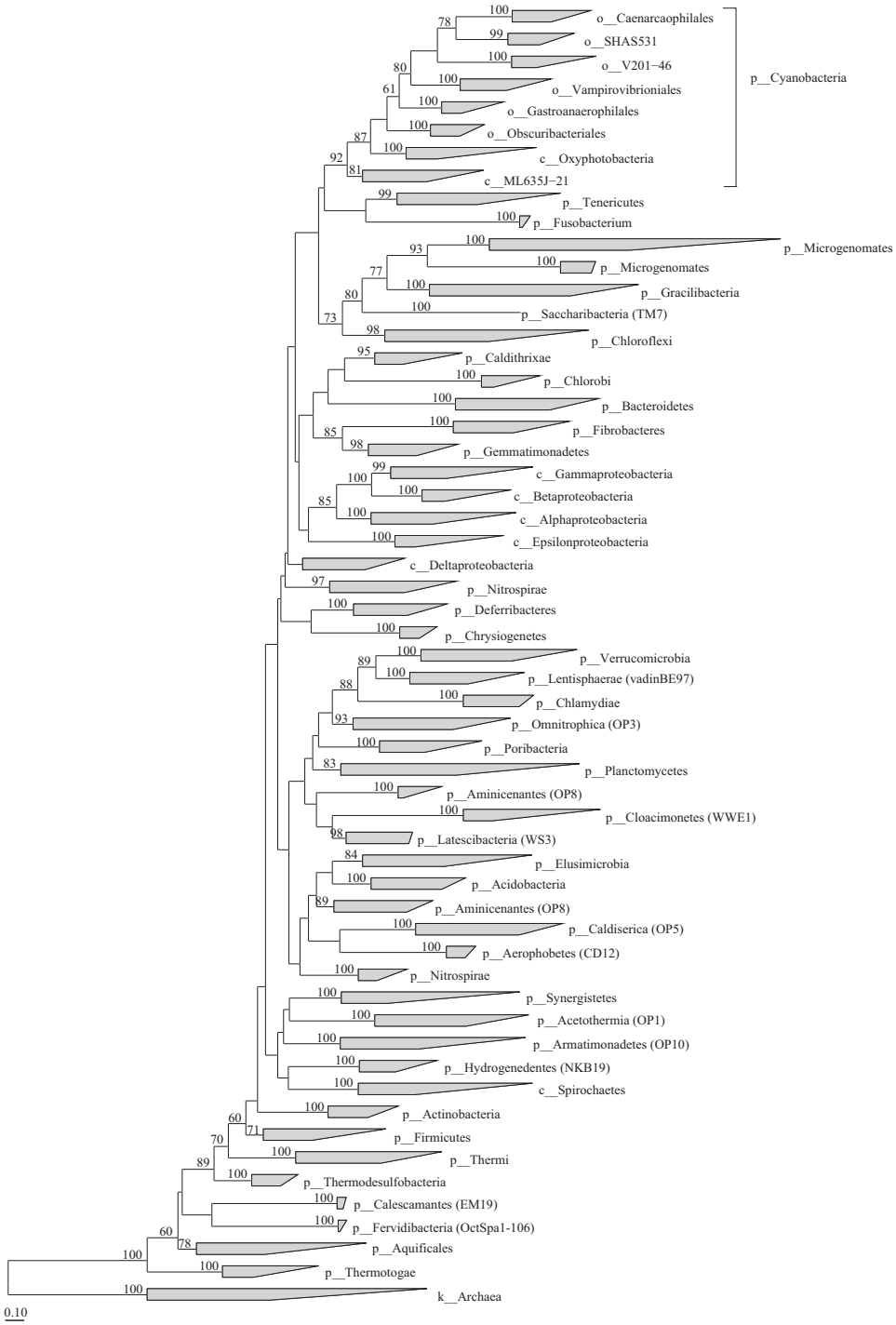


Figure S2.4. 16S rRNA gene tree showing the phylogenetic robstness of the Cyanobacteria and its class level lineages in the bacterial domain

Maximum likelihood (RAxML, GTR, G, I; FastTree, JTT, CAT) trees based on the 16S rRNA gene, showing the phylogenetic robustness of the classes Oxyphotobacteria and Melainabacteria. 418 OTUs (operational taxonomic units) from Bacteria and Archaea were used to construct the trees. Bootstrap analyses (100 times) were performed for the data set with maximum likelihood (RAxML, GTR, G, I; FastTree, JTT, CAT). Genomes in green are representatives from Di Rienzi *et al.,* 2013, genomes in red are Melainabacteria from this study and blue is the cultured representative, *Vampirovibrio chlorellavorus*. Bootstrap values >60% are shown *Candidatus* has been abbreviated to *Ca.* for the most complete genomes.

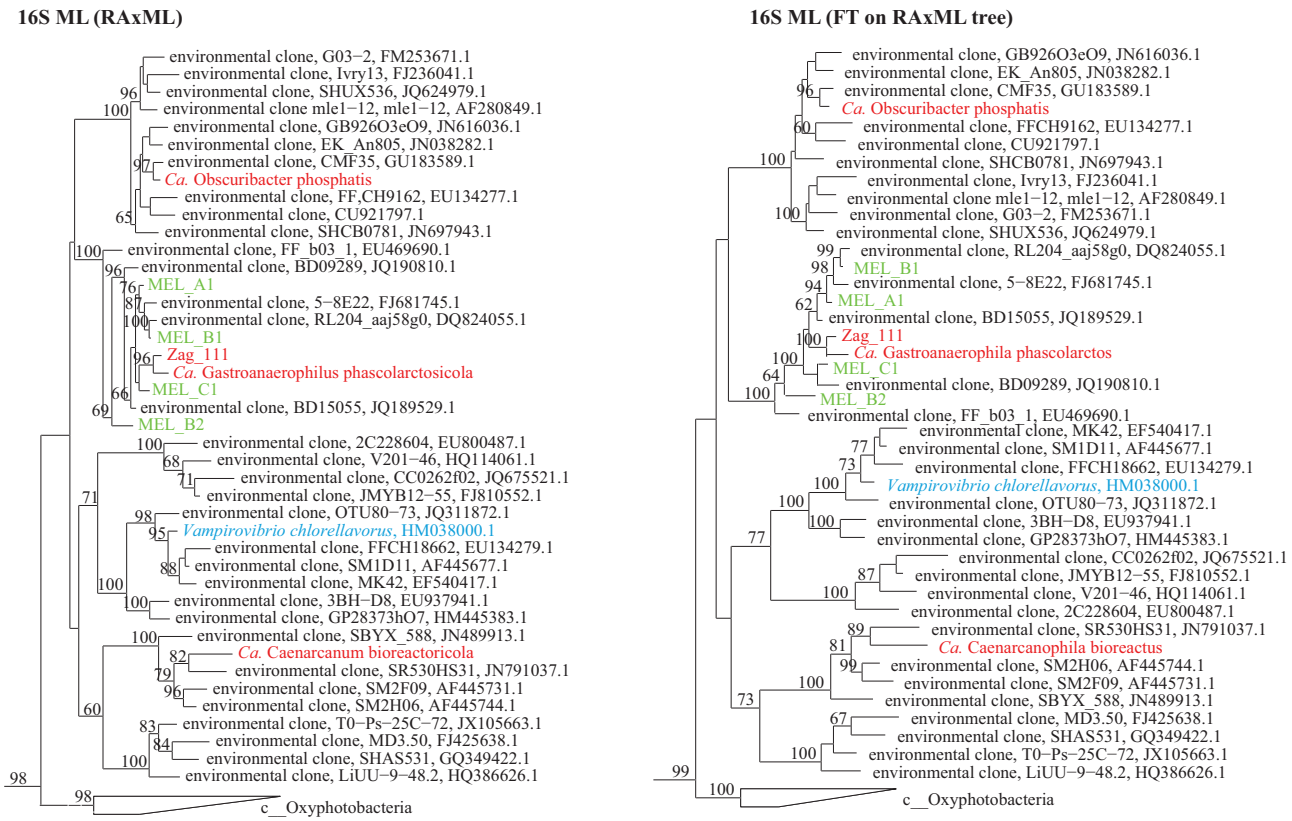


Figure S2.5. 16S rRNA gene tree showing the phylogenetic robustness of the order level taxa in the Melainabacteria

Phylogenetic neighbor joining (PAUP*, LogDet) and maximum parsimony (PAUP*) trees based on 16S rRNA gene, showing the phylogenetic robustness of the classes Oxyphotobacteria and Melainabacteria. 418 OTUs (operational taxonomic units) from Bacteria and Archaea were used to construct the trees. Bootstrap analyses (100 times) were performed for the data set with neighbor joining (PAUP*, LogDet) and maximum parsimony (PAUP*) methods. Genomes in green are representatives from Di Rienzi *et al.,* 2013, genomes in red are Melainabacteria from this study and blue is the cultured representative, *Vampirovibrio chlorellavorus*. Bootstrap values >60% are shown *Candidatus* has been abbreviated to *Ca.* for the most complete genomes.

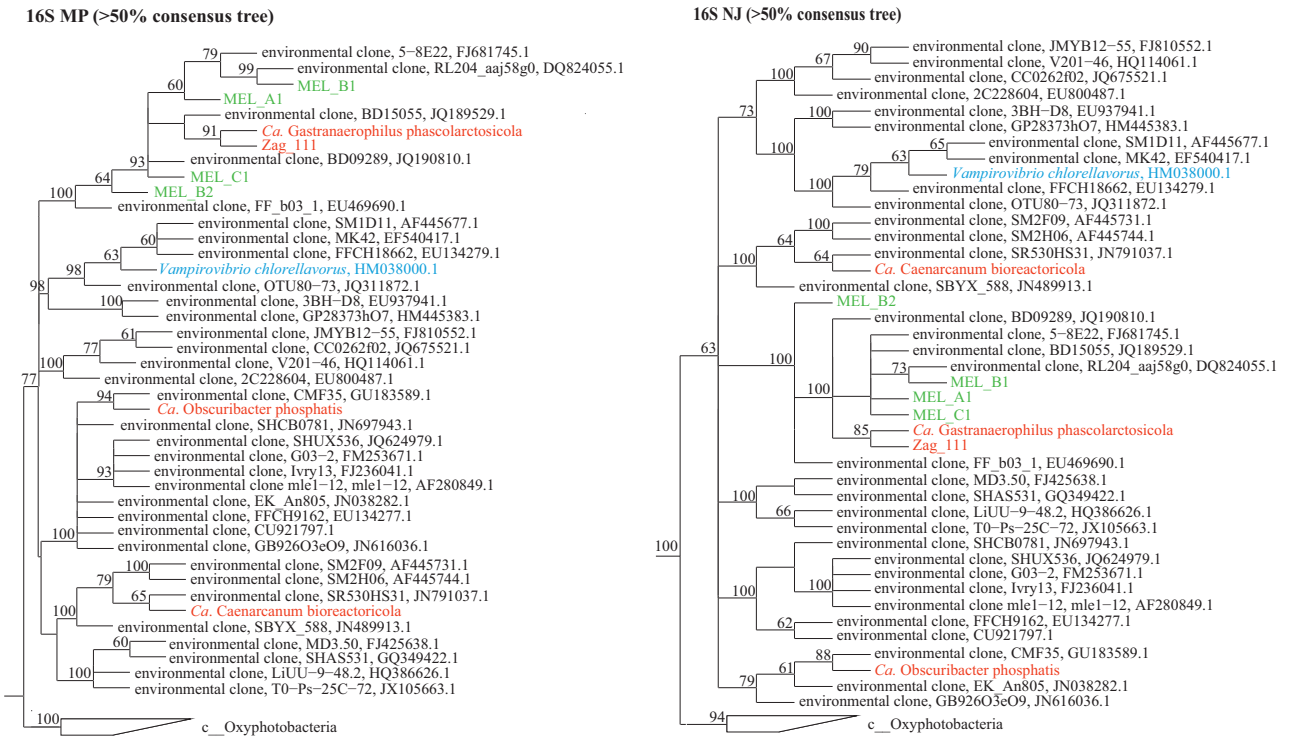


Figure S2.6. 16S rRNA gene tree showing the phylogenetic robustness of the order level taxa in the Melainabacteria

Phylogenetic neighbor joining (PAUP*, LogDet) and maximum parsimony (PAUP*) trees based on 16S rRNA gene, showing the phylogenetic robustness of the classes Oxyphotobacteria and Melainabacteria. 418 OTUs from Bacteria and Archaea were used to produce the trees. Bootstrap analyses (100 times) were performed for the data set with neighbor joining (PAUP*, LogDet) and maximum parsimony (PAUP*) methods, and the values obtained are shown in respective trees. 16S representatives in green are Melainabacterial representatives from [12] and the 16S from the cultured representative, *Vampirovibrio chlorellavorus* is in blue. 16S representatives in red are Melainabacteria from the current study. Genomes in green are representatives from Di Rienzi *et al.,* 2013, genomes in red are Melainabacteria from this study and blue is the cultured representative, *Vampirovibrio chlorellavorus*. Bootstrap values >60% are shown *Candidatus* has been abbreviated to *Ca.* for the most complete genomes.

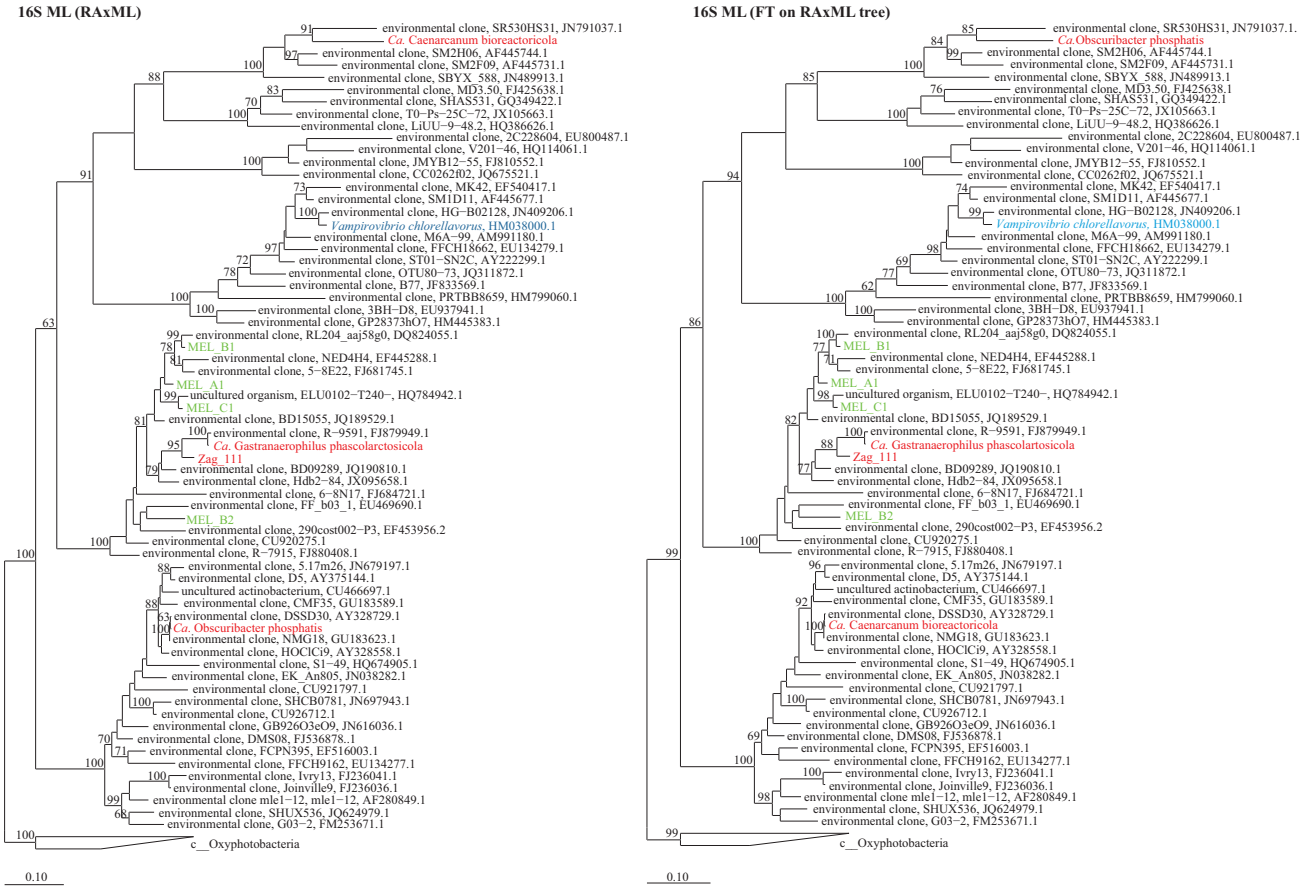


Figure S2.7. 16S rRNA gene tree showing the phylogenetic robustness of the order level taxa in the Melainabacteria

Phylogenetic maximum likelihood (RAxML, GTR, G, I; FastTree, JC, CAT) trees based on 16S rRNA gene, showing the phylogenetic robustness of order level lineages in the Melainabacteria. 67 OTUs from Cyanobacteria were used as outgroups. Bootstrap analyses (100 times) were performed for the data set with maximum likelihood methods (RAxML, GTR, G, I; FastTree, JC, CAT) and the values obtained are shown in respective trees, except for the values from FastTree, which are shown on the tree obtained from RAxML. 16S representatives in green are Melainabacterial representatives from Di Rienzi *et al.,* 2013 and the 16S from the cultured representative, *Vampirovibrio chlorellavorus* is in blue. 16S representatives in red are Melainabacteria from the current study. Bootstrap values >60% are shown. *Candidatus* has been abbreviated to *Ca.* for the most complete genomes.

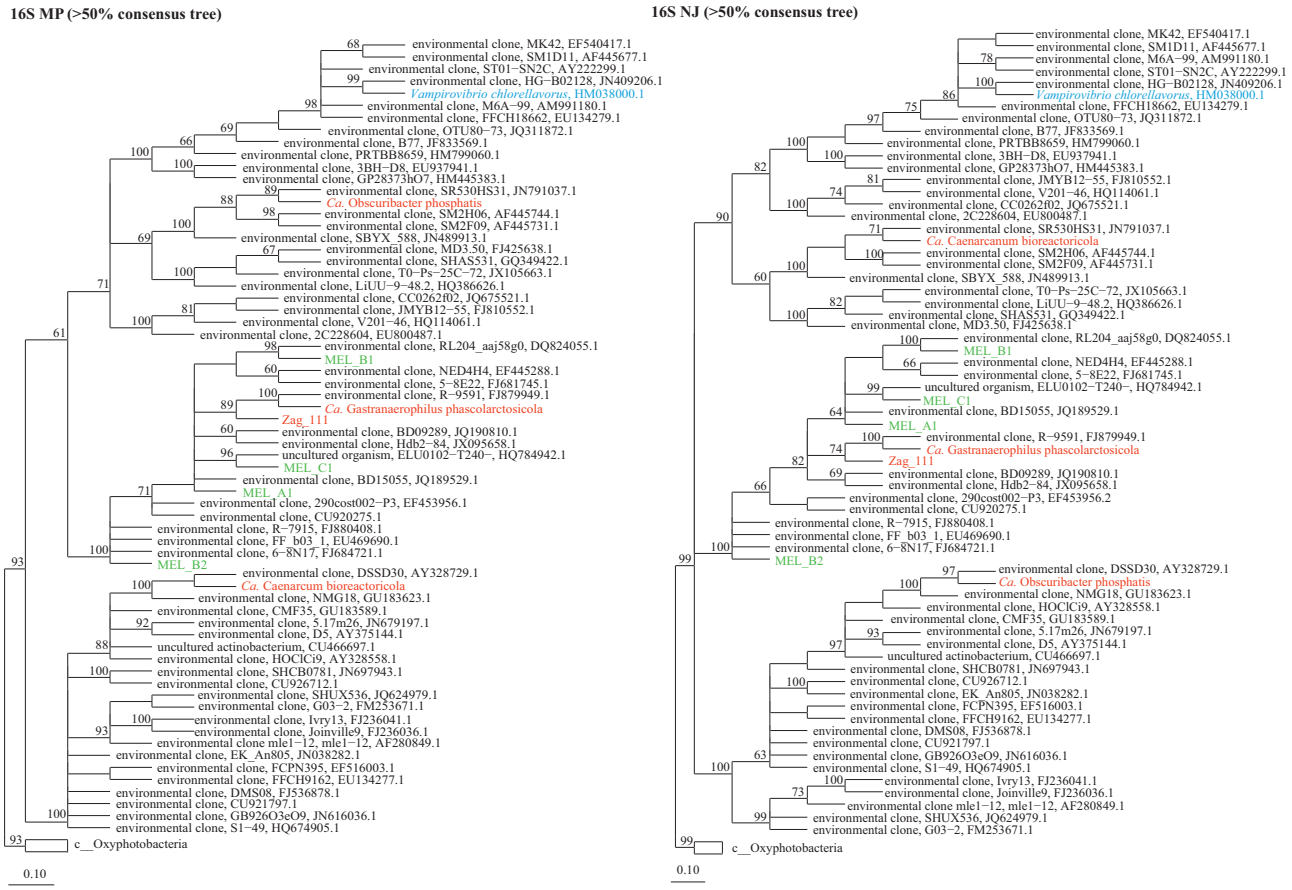


Figure S2.8. 16S rRNA gene tree showing the phylogenetic robustness of the order level taxa in the Melainabacteria

Phylogenetic maximum parsimony (PAUP*) and neighbor joining (PAUP*, LogDet) and trees based on 16S rRNA gene, showing the phylogenetic robustness of order level lineages in the Melainabacteria. 67 OTUs (operational taxonomic units) from Cyanobacteria were used as outgroups. Bootstrap analyses (100 times) were performed for the data set with maximum parsimony (PAUP*) and neighbor joining (PAUP*, LogDet) methods, and the values obtained are shown in >50% boostrap consensus trees. 16S representatives in green are Melainabacteria representatives from Di Rienzi *et al.,* 2013, and the 16S from the cultured representative, *Vampirovibrio chlorellavorus* is in blue. 16S representatives in red are Melainabacteria from the current study. Bootstrap values >60% are shown. *Candidatus* has been abbreviated to *Ca.* for the most complete genomes.

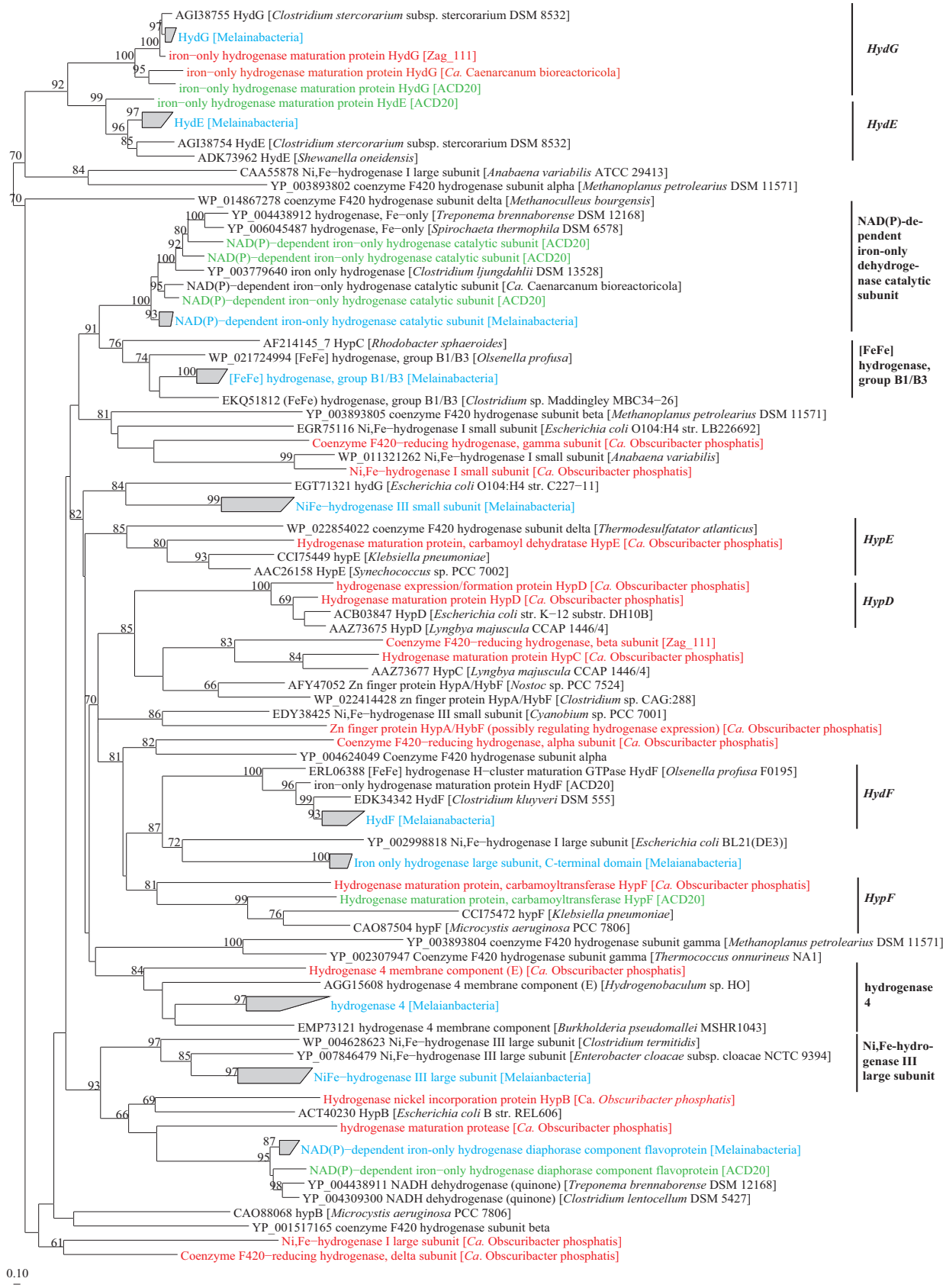


Figure S2.9. Hydrogenase gene tree

Gene tree showing the hydrogenase genes found in the Melainabacteria representatives. Known hydrogenase genes from NCBI were used to identify the phylogenetic positions for the hydrogenase genes.

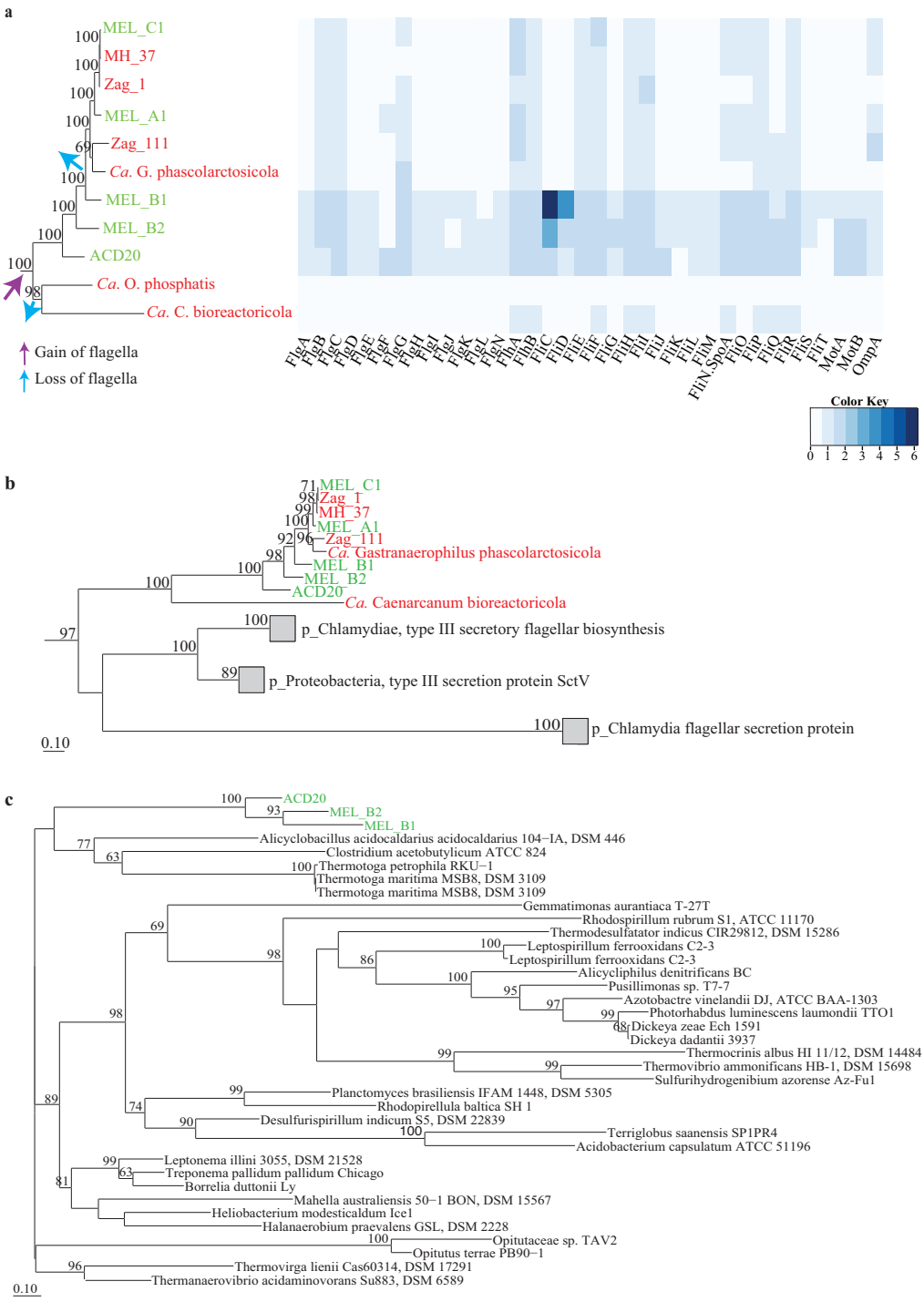


Figure S2.10. Heat map showing the presence and absence of flagella genes from the Melainabacteria representatives

(a) A maximum parsimony tree based on a concatenated alignment of up to 83 marker genes. Purple arrows indicate the gain of a functional flagella and blue arrows indicate the loss of a functional flagella. The heat map on the right indicates the square root of the number of flagella genes found in each of the Melainabacteria representatives. (b) Flagella gene trees for FlhA and (c) flagella gene tree for FliF. Genomes used to make the trees are found in **Table S2.6**. Genomes in green are Melainabacteria representatives from Di Rienzi *et al.,* 2013, genomes in red are Melaianabacteria representatives from this study. *Candidatus* has been abbreviated to *Ca.* for the most complete genomes.

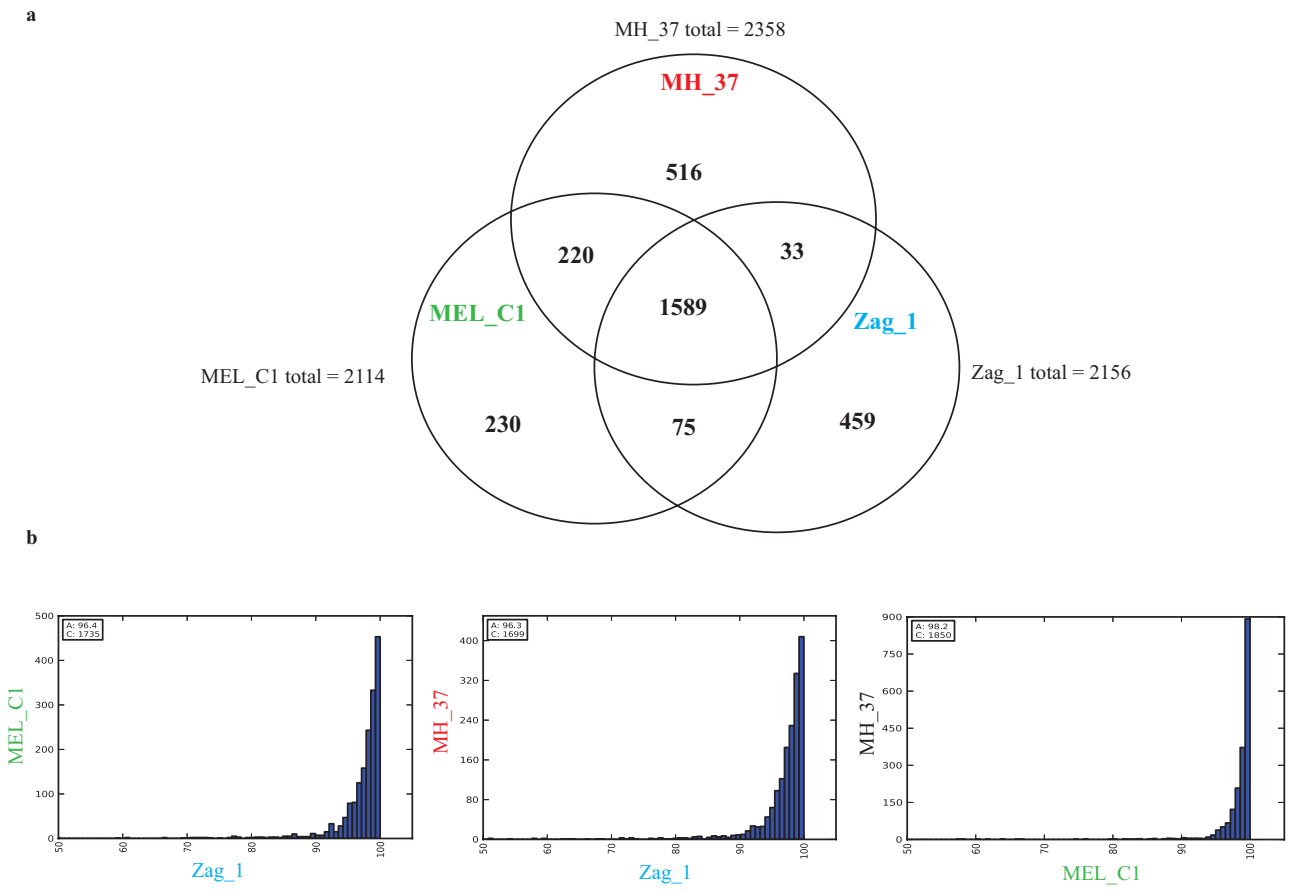


Figure S2.11. Venn diagram of the three Gastranaerophilales genomes from the same species
(a) Venn diagram between MH_37, MEL_C1 and Zag_1 showing the number of genes that are core to the three genomes, and those that are found between two genomes and individual genomes. The total number for each genomes does not include paralogs. (b) Average nucleotide identity between MH_37, MEL_C1 and Zag_1, where ≥95% ANI is used to define a species.

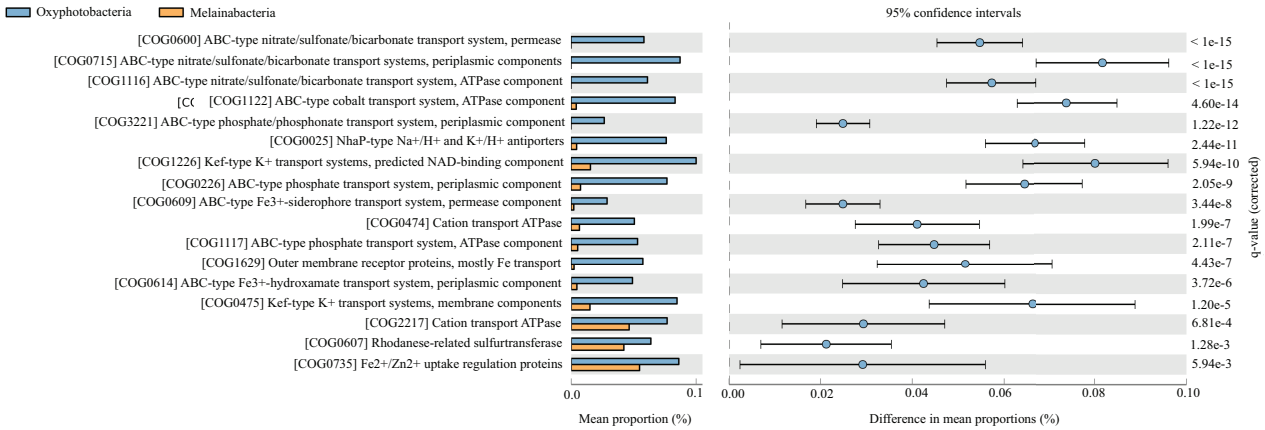


Figure S2.12. COG category P (inorganic ion transport and metabolism) for Oxyphotobacteria and Melainabacteria

The category specific plots show all COGs that differ by at least 0.02% and have a Storey q-value of < 0.05 between Oxyphotobacteria and Melainabacteria, where the blue bars represent all complete Oxyphotobacteria genomes from IMG and the orange bars represent Melainabacteria.

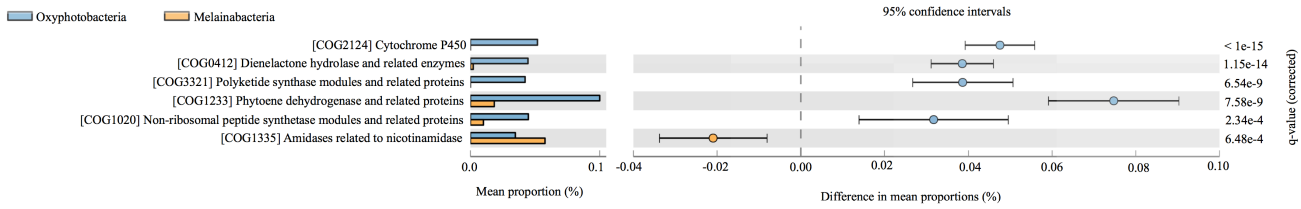


Figure S2.13. COG category Q (secondary metabolites and biosynthesis, transport and catabolism) for Oxyphotobacteria and Melaianabacteria

The category specific plots show all COGs that differ by at least 0.02% and have a Storey q-value of < 0.05 between Oxyphotobacteria and Melainabacteria, where the blue bars represent all complete Oxyphotobacteria genomes from IMG and the orange bars represent Melainabacteria.

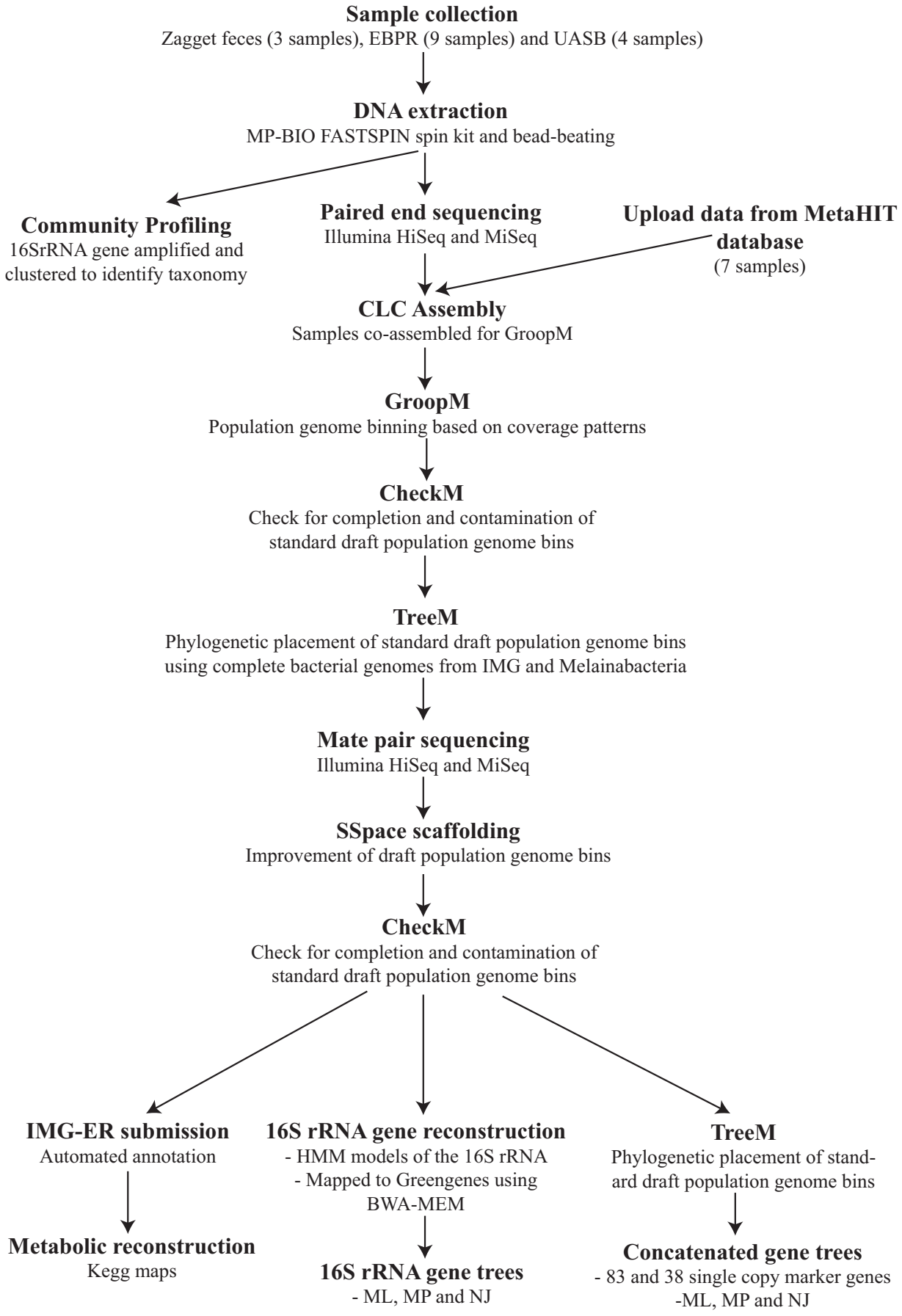


Figure S2.14. Flow diagram of the Methods used in this paper

Supplementary Tables

Table S2.1. Sequencing statistics

EBPR1_T1 to EBPR1_T6 and EBPR2_T1 to EBPR2_T3 (blue) correspond to the nine samples collected from two enhanced biological phosphorous removal bioreactor (EBPR), Zag_T1 to Zag_T3 (red) correspond to the three time points where samples were collected from koala feces, MH_F2, F3, F5, F6, F8, M3 and M8 correspond to human feces collected from seven Danish females and males (http://www.metahit.eu/) (purple) and A1, A2, F1 and F2 from the UASB (green). The combined assembly statistics is the amount of sequencing performed for all of the samples collected from EBPR, koala feces, Danish individuals or UASB. Genome population bins is the number of genome bins that was produced by GroopM v.1.0. The N50 is for all of the combined metagenomic data for each sample.

		Combined shotgun sequencing assembly statistics for GroopM				Mate pair sequencing	
Sample ID	**Sampling Date**	**Shotgun sequence (Gbp)**	**Number of contigs**	**N50**	**Genome population bins**	**Mate pair sequencing (Gbp)**	**Insert size (kbp)**
EBPR1_T1	05/27/11	26.2 (174,720,232 x 150bp)	148,338	1.4 kbp	299	4.88	3.2-3.8
EBPR1_T2	06/22/11	21.49 (143,317,626 x 150bp)					
EBPR1_T3	08/01/11	23.76 (158,451,996 x 150bp)					
EBPR1_T4	09/08/11	38.01 (253,461,380 x 150bp)					
EBPR1_T5	11/25/11	16.78 (111,886,692 x 150bp)					
EBPR1_T6	01/18/12	22.65 (151,023,810 x 150bp)				5.72	3.2-3.8
EBPR2_T1	06/17/11	18.86 (125,756,890 x 150bp)				6.15	3.2-3.8
EBPR2_T2	09/05/11	19.08 (127,217,526 x 150bp)					
EBPR2_T3	12/16/11	24.81 (165,419,806 x 150bp)				5.10	3.2-3.8
Zag_T1	05/12/11	10.68 (71,196,258 x 150bp)	17,101	4.6 kbp	181	1.99	2-15
Zag_T2	07/28/11	65.06 (433,763,472 x 150bp)				1.89	2-15
Zag_T3	11/24/11	15.00 (100,022,578 x 150bp)				1.87	2-15
UASB_A1	12/25/12	7.3 (32,365,294 x 250bp)	84,262	2.7 kbp	154	2.00	5.5
UASB_A2	09/16/10	5.8 (23,420,132 x 250bp)					
UASB_F1	-	5.0 (23,027,350 x 250bp)					
UASB_G1	-	7.8 (35,467,130 x 250bp)					

MH_F2 (M0002)	-	3.49 (46,574,230 x 75bp)	3,139	2.3 kbp	119		
MH_F3 (MH0006)	-	6.96 (92,764,998 x 75bp)					
MH_F5 (MH0021)	-	1.97 (26,262,536 x 75bp)					
MH_F6 (MH0024)	-	1.61 (21,421,864 x 75bp)					
MH_F8 (MH0028)	-	1.55 (20,637,196 x 75bp)					
MH_M3 (MH0009)	-	4.38 (58,458,876 x 75bp)					
MH_M8 (MH0031)	-	1.62 (21,566,850 x 75bp)					

Table S2.2. List of 83 single copy gene markers

The list of 83 single copy gene markers is a subset of the 111 single copy gene markers compiled by Dupont *et al.,* 2012.

TIGR/PFAM	Name	Size (aa)
TIGR00064	ftsY: Signal recognition particle-docking protein FtsY	279
TIGR00082	rbfA: ribosome-binding factor A	115
TIGR00086	smpB: SsrA-binding protein	144
TIRG00092	TIGR00092: GTP-binding protein YchF	368
TIGR00115	tig: trigger factor	410
TIGR00116	tsf: translation elongation factor Ts	293
TIGR00158	L9: ribosomal protein L9	148
TIGR00165	S18: ribosomal protein S18	70
TIGR00337	PyrG: CTP synthase	526
TIGR00344	alaS: alanine—tRNA ligase	847
TIGR02386	rpoC_TIGR: DNA-directed RNA polymerase, beta subunit	1147
TIGR02387	rpoC1_cyan: DNA-directed RNA polymerase, gamma subunit	619
TIGR02397	dnaX_nterm: DNA polymerase III, subunit gamma and tau	355
TIGR02729	Obg_CgtA: Obg family GTPase CgtA	329
TIGR02012	tigrfam_recA: protein RecA	321
TIGR02013	rpoB: DNA-directed RNA polymerase, beta subunit	1238
TIGR02027	rpoA: DNA-directed RNA polymerase, alpha subunit	298
TIGR03263	guanyl_kin: guanylate kinase	180
TIGR03594	GTPase_EngA: ribosome-associated GTPase EngA	432
TIGR00409	proS_fam_II: proline—tRNA ligase	568
PF00162	Phosphoglycerate kinase	384
PF00276	Ribosomal protein L23	92
PF00281	Ribosomal protein L5	56
PF00297	Ribosomal protein L3	263
PF00380	Ribosomal protein S9/S16	121
PF00410	Ribosomal protein S8	129
PF00411	Ribosomal protein S11	110
PF00416	Ribosomal protein S13/S18	106
PF00466	Ribosomal protein L10	100
PF00573	Ribosomal protein L4/L1 family	192
PF01795	MraW methylase family	310
TIGR00001	rpmI_bact: ribosomal protein L35	63
TIGR00002	S16: ribosomal protein S16	78
TIGR00019	prfA: peptide chain release factor 1	361
TIGR00029	S20: ribosomal protein S20	87
TIGR00043	TIGR00043: metalloprotein, YbeY/UPF0054 family	111
TIGR00059	L17: ribosomal protein L17	112
TIGR00060	L18_bact: ribosomal protein L18	114
TIGR00061	L21: ribosomal protein L21	101
TIGR00166	S6: ribosomal protein S6	95
TIGR00168	infC: translation initiation factor IF-3	165
TIGR00362	DnaA: chromosomal replication initiator protein DnaA	437
TIGR00388	glyQ: glycine—tRNA ligase, alpha subunit	293
TIGR00389	glyS_dimeric: glycine—tRNA ligase	565
TIGR00459	aspS_bact: aspartate—tRNA ligase	586

TIGR00460	fmt: methionyl-tRNA formyltransferase	315
TIGR00468	pheS: phenylalainin—tRNA ligase, alpha subunit	324
TIGR00959	ffh: signal recognition particle protein	428
TIGR00963	secA: preprotein translocase, SecA subunit	787
TIGR00964	secE_bact: preprotein translocase, SecE subunit	57
TIGR00967	3a0501s007: preprotein translocase, SecY subunit	414
TIGR00981	rpsL_bact: ribosomal protein S12	124
TIGR01009	rpsC_bact: ribosomal protein S3	212
TIGR01011	rpsB_bact: ribosomal protein S2	225
TIGR01021	rpsE_bact: ribosomal protein S5	156
TIGR01024	rplS_bact: ribosomal protein L19	114
TIGR01029	rpsG_bact: ribosomal protein S7	154
TIGR01032	rplT_bact: ribosomal protein L20	114
TIGR01044	rplV_bact: ribosomal protein L22	103
TIGR01049	rpsJ_bact: ribosomal protein S10	99
TIGR01050	rpsS_bact: ribosomal protein S19	92
TIGR01063	gyrA: DNA gyrase, A subunit	800
TIGR01066	rplM_bact: ribosomal protein L13	141
TIGR01067	rplN_bact: ribosomal protein L14	122
TIGR01071	rplO_bact: ribosomal protein L15	144
TIGR01079	rplX_bact: ribosomal protein L24	104
TIGR00471	pheT_arch: phenylalanine—tRNA ligase, beta subunit	551
TIGR00472	pheT_bact: phenylalanine—tRNA ligase, beta subunit	798
TIGR00487	IF-2: translation initiation factor IF-2	587
TIGR00496	frr: ribosome recycling factor	176
TIGR00575	dnlj: DNA ligase, NAD-dependent	652
TIGR00631	uvrb: excinuclease ABC subunit B	658
TIGR00663	dnan: DNA polymerase III, beta subunit	367
TIGR00810	secG: preprotein translocase, SecG subunit	73
TIGR00855	L12: ribosomal protein L7/L12	125
TIGR00922	nusG: transcription termination/antitermination factor NusG	172
TIGR01164	rplP_bact: ribosomal protein L16	126
TIGR01169	rplA_bact: ribosomal protein L1	227
TIGR01171	rplB_bact: ribosomal protein L2	275
TIGR01391	dnaG: DNA primase	414
TIGR01393	lepA: GTP-binding protein LepA	595
TIGR01632	L11_bact: ribosomal protein L11	140
TIGR01953	NusA: transcription termination factor NusA	340

Table S2.3. List of Oxyphotobacteria, Melainabacteria and Chloroflexi Oxyphotobacteria representatives from IMG, as grouped by Shih *et al.,* 2013 (group A to group G), Melainabacteria from Di Rienzi *et al.,* 2013 and this study, as well as Chloroflexi from JGI IMG that were used to make the concatenated gene tree (**Figure 2.1A)**.

Phylum Cyanobacteria	
Class Oxyphotobacteria	**IMG Accession**
Group A	
Arthrospira maxima CS-328	642979357
Arthrospira platensis C1	2507262036
Arthrospira sp. PCC 8005	648276619
Arthrospira platensis NIES-39	650377906
Lyngbya sp. CCY 8106	639857035
Trichodesmium erythraeum IMS101	637000329
Oscillatoria sp. PCC 6506	648276706
Oscillatoria nigro-viridis PCC 7112	2503982035
Oscillatoria acuminata PCC 6304	2509276028
Group B	
Cyanothece sp. BH68, ATCC 51142	641522622
Crocosphaera watsonii WH 8501	2531839001
Cyanobacterium UCYN-A	646311970
Cyanothece sp. PCC 8801	643348535
Synechocystis sp. PCC 6803	637000315
Cyanothece sp. PCC 7424	643348533
Cyanothece sp. PCC 7822	648028021
Microcystis aeruginosa NIES-843	641522640
Pleurocapsa sp. PCC 7327	2509276061
Synechococcus sp. PCC 7002	641522654
Leptolyngbya sp. PCC 7376	2503754048
Cyanobacterium stanieri PCC 7202	2503283023
Cyanobacterium aponinum PCC 10605	2503707009
Stanieria cyanosphaera PCC 7437	2503754019
Chroococcidiopsis sp. PCC 6712	2505679029
Spirulina major PCC 6313	2506520014
Spirulina subsalsa PCC 9445	2506520011
Dactylococcopsis salina PCC 8305	2509276056
Halothece sp. PCC 7418	2503538028
Microcoleus chthonoplastes PCC 7420	647533184
Microcoleus sp. PCC 7113	2509276031
Cylindrospermopsis raciborskii CS-505	647000233
Raphidiopsis brookii D9	647000303
Nostoc azollae 0708	648028001
Anabaena cylindrica PCC 7122	2503982047
Anabaena sp. PCC 7108	2506485002
Nostoc punctiforme PCC 73102	642555144
Calothrix sp. PCC 7507	2505679032
Nodularia spumigena CCY9414	639857037
Nostoc sp. PCC 7120	637000199
Anabaena variabilis ATCC 29413	646564504

Nostoc sp. PCC 7524	2509601032
Nostoc sp. PCC 7107	2503707008
Rivularia sp. PCC 7116	2510065008
Calothrix sp. PCC 6303	2503982036
Fischerella sp. JSC-11	2505679024
Gloeocapsa sp. PCC 7428	2503754017
Chroococcidiopsis thermalis PCC 7203	2503538021
Crinalium epipsammum PCC 9333	2504643013
Group C	
Synechococcus sp. CC9616	2517093019
Prochlorococcus sp. WH8102	637000314
Prochlorococcus sp. CC9605	637000310
Prochlorococcus sp. CC9902	637000311
Prochlorococcus sp. CC9311	637000309
Synechococcus sp. WH 8016	2507262052
Prochlorococcus sp. WH 7803	640427149
Prochlorococcus marinus pastoris CCMP 1986	637000214
Prochlorococcus marinus MIT 9515	640069324
Prochlorococcus marinus AS9601	640069321
Prochlorococcus marinus NATL2A	637000212
Prochlorococcus marinus marinus CCMP 1375	637000213
Prochlorococcus marinus MIT 9211	641228501
Prochlorococcus marinus MIT 9313	637000211
Cyanobium sp. PCC 7001	647533126
Cyanobium gracile PCC 6307	2508501011
Synechococcus sp. RCC307	640427148
Synechococcus elongatus PCC 6301	637000307
Synechococcus elongatus PCC 7942	637000308
Group D	
Cyanobacterium sp. JSC-1	2502171143
Oscillatoriales sp. JSC-12	2510065010
Group E	
Thermosynechococcus elongatus BP-1	637000320
Synechococcus sp. PCC 6312	2509276030
Cyanothece sp. PCC 7425	643348534
Acaryochloris marina MBIC11017	641228474
Group F	
Pseudanabaena sp. PCC 7367	2504643012
Synechococcus sp. PCC 7502	2508501041
Group G	
Synechococcus sp. JA-3-3Ab	637000313
Synechococcus sp. PE A4 65AY6A	2512875021
Synechococcus sp. JA-2-3B	637000312
Synechococcus sp. PCC 7336	2506520048
Gloeobacter violaceus PCC 7421	637000121
Geitlerinema sp. PCC 7407	2503538020
Class Melainabacteria	
Ca. Obscuribacter phosphatis	2541046960

Ca. Caenarcanophila bioreactus	2531839742
Ca. Gastroanaerophila phascolarctos	2523533519
Za_1	2523533517
Zag_111	2531839741
MH_37	2522572068
ACD20	2541046958
MEL_A1	2541016959
MEL_B1	2541046956
MEL_B2	2541046940
MEL_C1	2541046938
Phylum Chloroflexi	
Class Chloroflexi	
Chloroflexus aurantiacus J-10-fl	641228485
Chloroflexus aggregans DSM 9485	643348527
Oscillochloris trichoides DG6	649989977
Roseiflexus sp. RS-1	640427139
Roseiflexus castenholzii HLO8, DSM 13941	640753047
Herpetosiphon aurantiacus DSM 785	2508501111
Class Thermomicrobia	
Thermomicrobium roseum DSM 5159	643348582
Sphaerobacter terhmophilus 4ac11, DSM 20745	646311953
Thermobaculum terrenum YNP1 ATCC BAA-798	646311962
Class Anaerolineae	
Anaerolinea thermophila UNI-1	649633005
Class Dehalococcoidia	
Dehalococcoides sp. BAV1	640427111
Dehalogenimonas lykanthroporepellens BL-DC-9	648028022

Table S2.4. List of Oxyphotobacteria, Melainabacteria and outgroups using universal single copy bacterial marker gene sets

Bacterial and archaeal genomes used to produce **Figures S2.3** and **S2.4**. Organisms in black were used to produce phylogenetic trees using both the 38 marker and 83 marker sets. Organisms in blue were used for phylogenetic trees made with the 83 marker set only and organisms in red were used for phylogenetic trees made with the 38 marker set.

Phylum	Organism name	IMG/NCBI accession
Acidobacteria	*Acidobacterium capsulatum* ATCC 51196	643692001
Acidobacteria	*Granulicella mallensis* MP5ACTX8	648276601
Acidobacteria	*Korebacter versatilis* Ellin345	637000001
Acidobacteria	*Solibacter usitatus* Ellin6076	639633060
Acidobacteria	*Terriglobus saanensis* SP1PR4	649633100
Actinobacteria	*Acidimicrobium ferrooxidans* ICP, DSM 10331	644736322
Actinobacteria	*Atopobium parvulum* IPP 1246, DSM 20469	644736327
Actinobacteria	*Bifidobacterium longum* DJO10A	642555107
Actinobacteria	*Conexibacter woesei* ID131577, DSM 14684	646311917
Actinobacteria	*Corynebacterium efficiens* YS-314	644736345
Actinobacteria	*Cryptobacterium curtum* 12-3, DSM 15641	644736346
Actinobacteria	*Eggerthella lenta* VPI 0255, DSM 2243	644736358
Actinobacteria	*Gordonibacter pamelaeae* 7-10-1-bT, DSM 19378	650377943
Actinobacteria	*Leifsonia xyli xyli* CTCB07	637000149
Actinobacteria	*Micrococcus luteus* Fleming NCTC 2665	644736390
Actinobacteria	*Microlunatus phosphovorus* NM-1	650716058
Actinobacteria	*Micromonospora* sp. L5	649633069
Actinobacteria	*Propionibacterium freudenreichii shermanii* CIRM-BIA1	649633084
Actinobacteria	*Rhodococcus jostii* RHA1	637000234
Actinobacteria	*Rubrobacter xylanophilus* DSM 9941	637000248
Actinobacteria	*Streptomyces scabiei* 87.22	646564576
Actinobacteria	*Thermobifida fusca* YX	637000319
Aquificae	*Aquifex aeolicus* VF5	637000010
Aquificae	*Desulfurobacterium thermolithotrophum* BSA, DSM 11699	649633039
Aquificae	*Hydrogenivirga* sp. 128-5-R1-1	641380441
Aquificae	*Hydrogenobacter thermophilus* TK-6, DSM 6534	646311936
Aquificae	*Hydrogenobaculum* sp. SN	647000261
Aquificae	*Persephonella marina* EX-H1	643692030
Aquificae	*Sulfurihydrogenibium azorense* Az-Fu1	643692050
Aquificae	*Sulfurihydrogenibium* sp. YO3AOP1	642555165
Aquificae	*Thermocrinis albus* HI 11/12, DSM 14484	646564582
Aquificae	*Thermovibrio ammonificans* HB-1, DSM 15698	649633104
Bacteroidetes	*Bacteroides fragilis* 3_1_12	645058788
Bacteroidetes	*Bacteroidetes* sp. F0058	648861005

Bacteroidetes	*Capnocytophaga gingivalis* ATCC 33624	643886113
Bacteroidetes	*Chitinophaga pinensis* UQM 2034, DSM 2588	644736340
Bacteroidetes	*Chryseobacterium gleum* F93, ATCC 35910	643886082
Bacteroidetes	*Croceibacter atlanticus* HTCC2559	648028020
Bacteroidetes	*Cytophaga hutchinsonii* ATCC 33406	637000087
Bacteroidetes	*Flavobacterium johnsoniae* UW101, ATCC 17061	644736369
Bacteroidetes	*Kordia algicida* OT-1	641380434
Bacteroidetes	*Leadbetterella byssophila* 4M15, DSM 17132	649633063
Bacteroidetes	*Parabacteroides merdae* ATCC 43184	640963016
Bacteroidetes	*Porphyromonas asaccharolytica* PR426713P-I	649989985
Bacteroidetes	*Porphyromonas endodontalis* ATCC 35406	643886148
Bacteroidetes	*Porphyromonas gingivalis* ATCC 33277	642555148
Bacteroidetes	*Prevotella melaninogenica* ATCC 25845	648028051
Bacteroidetes	*Prevotella tannerae* ATCC 51259	645951840
Bacteroidetes	*Psychroflexus torquis* ATCC 700755	638341165
Bacteroidetes	*Sphingobacterium spiritivorum* ATCC 33300	643886135
Chlamydiae	*Chlamydia muridarum* MoPn / Nigg	637000062
Chlamydiae	*Chlamydia trachomatis* A/HAR-13	637000063
Chlamydiae	*Chlamydophila abortus* S26/3	637000065
Chlamydiae	*Chlamydophila caviae* GPIC	637000066
Chlamydiae	*Chlamydophila felis* Fe/C-56	637000067
Chlamydiae	*Chlamydophila pecorum* E58	650716022
Chlamydiae	*Chlamydophila pneumoniae* AR39	637000068
Chlamydiae	*Chlamydophila psittaci* 01DC11	651053012
Chlamydiae	*Parachlamydia acanthamoebae* Hall's coccus	647000287
Chlamydiae	*Simkania negevensis* Z	650716085
Chlamydiae	*Waddlia chondrophila* WSU 86-1044	646564588
Chlorobi	*Chlorobaculum parvum* NCIB 8327	642555120
Chlorobi	*Chlorobium chlorochromatii* CaD3	637000072
Chlorobi	*Chlorobium limicola* DSM 245	642555121
Chlorobi	*Chlorobium phaeobacteroides* BS1	642555122
Chlorobi	*Chlorobium phaeovibrioides* DSM 265	640427130
Chlorobi	*Chlorobium tepidum* TLS	637000073
Chlorobi	*Chloroherpeton thalassium* ATCC 35110	642555123
Chlorobi	*Pelodictyon luteolum* DSM 273	637000205
Chlorobi	*Pelodictyon phaeoclathratiforme* BU-1	642555146
Chlorobi	*Prosthecochloris aestuarii* SK413, DSM 271	642555149
Chloroflexi	*Anaerolinea thermophila* UNI-1	649633005
Chloroflexi	*Chloroflexus aggregans* DSM 9485	643348527
Chloroflexi	*Chloroflexus aurantiacus* J-10-fl	641228485
Chloroflexi	*Dehalococcoides ethenogenes* 195	637000089
Chloroflexi	*Dehalococcoides* sp. BAV1	640427111
Chloroflexi	*Dehalogenimonas lykanthroporepellens* BL-DC-9	648028022
Chloroflexi	*Herpetosiphon aurantiacus* DSM 785	641228494
Chloroflexi	*Oscillochloris trichoides* DG6	649989977
Chloroflexi	*Roseiflexus castenholzii* HLO8, DSM 13941	640753047

Chloroflexi	*Roseiflexus* sp. RS-1	640427139
Chloroflexi	*Sphaerobacter thermophilus* 4ac11, DSM 20745	646311953
Chloroflexi	*Thermobaculum terrenum* YNP1, ATCC BAA-798	646311962
Chloroflexi	*Thermomicrobium roseum* DSM 5159	643348582
Chrysiogenetes	*Desulfurispirillum indicum* S5, DSM 22839	649633038
Cyanobacteria	*Acaryochloris marina* MBIC11017	641228474
Cyanobacteria	ACD20	2541046958
Cyanobacteria	*Anabaena variabilis* ATCC 29413	646564504
Cyanobacteria	*Arthrospira maxima* CS-328	642979357
Cyanobacteria	*Arthrospira platensis* NIES-39	650377906
Cyanobacteria	*Arthrospira* sp. PCC 8005	648276619
Cyanobacteria	*Candidatus* Caenarcanophila bioreactus	2523533519
Cyanobacteria	*Candidatus* Gastroanaerophila phascolarctos	2523533519
Cyanobacteria	*Candidatus* Obscuribacter phosphatis	2541046960
Cyanobacteria	*Crocosphaera watsonii* WH 8501	638341074
Cyanobacteria	cyanobacterium UCYN-A	646311970
Cyanobacteria	*Cyanobium* sp. PCC 7001	647533126
Cyanobacteria	*Cyanothece* sp. BH68, ATCC 51142	641522622
Cyanobacteria	*Cyanothece* sp. PCC 7424	643348533
Cyanobacteria	*Cyanothece* sp. PCC 7425	643348534
Cyanobacteria	*Cyanothece* sp. PCC 7822	648028021
Cyanobacteria	*Cyanothece* sp. PCC 8801	643348535
Cyanobacteria	*Cylindrospermopsis raciborskii* CS-505	647000233
Cyanobacteria	*Gloeobacter violaceus* PCC 7421	637000121
Cyanobacteria	*Lyngbya* sp. CCY 8106	639857035
Cyanobacteria	MEL_A1	2541046959
Cyanobacteria	MEL_B1	2541046956
Cyanobacteria	MEL_B2	2541046940
Cyanobacteria	MEL_C1	2541046938
Cyanobacteria	MH_37	2522572068
Cyanobacteria	*Microcoleus chthonoplastes* PCC 7420	647533184
Cyanobacteria	*Microcystis aeruginosa* NIES-843	641522640
Cyanobacteria	*Nodularia spumigena* CCY9414	639857037
Cyanobacteria	*Nostoc azollae* 0708	648028001
Cyanobacteria	*Nostoc punctiforme* PCC 73102	642555144
Cyanobacteria	*Nostoc* sp. PCC 7120	637000199
Cyanobacteria	*Oscillatoria* sp. PCC 6506	648276706
Cyanobacteria	*Prochlorococcus marinus* AS9601	640069321
Cyanobacteria	*Prochlorococcus marinus marinus* CCMP1375	637000213
Cyanobacteria	*Prochlorococcus marinus* MIT 9211	641228501
Cyanobacteria	*Prochlorococcus marinus* MIT 9313	637000211
Cyanobacteria	*Prochlorococcus marinus* MIT 9515	640069324
Cyanobacteria	*Prochlorococcus marinus* NATL2A	637000212
Cyanobacteria	*Prochlorococcus marinus pastoris* CCMP1986	637000214
Cyanobacteria	*Prochlorococcus* sp. CC9311	637000309
Cyanobacteria	*Prochlorococcus* sp. CC9605	637000310

Cyanobacteria	*Prochlorococcus* sp. CC9902	637000311
Cyanobacteria	*Prochlorococcus* sp. WH 7803	640427149
Cyanobacteria	*Prochlorococcus* sp. WH8102	637000314
Cyanobacteria	*Raphidiopsis brookii* D9	647000303
Cyanobacteria	*Synechococcus elongatus* PCC 6301	637000307
Cyanobacteria	*Synechococcus* sp. CC9616	2514885022
Cyanobacteria	*Synechococcus* sp. CC9616	2517093019
Cyanobacteria	*Synechococcus* sp. JA-2-3B	637000312
Cyanobacteria	*Synechococcus* sp. JA-3-3Ab	637000313
Cyanobacteria	*Synechococcus* sp. PCC 7002	641522654
Cyanobacteria	*Synechococcus* sp. RCC307	640427148
Cyanobacteria	*Synechocystis* sp. PCC 6803	637000315
Cyanobacteria	*Thermosynechococcus elongatus* BP-1	637000320
Cyanobacteria	*Trichodesmium erythraeum* IMS101	637000329
Cyanobacteria	Zag_1	2523533517
Cyanobacteria	Zag_111	2531839741
Deferribacteres	*Calditerrivibrio nitroreducens* Yu37-1, DSM 19672	649633026
Deferribacteres	*Deferribacter desulfuricans* SSM1, DSM 14783	646564525
Deferribacteres	*Denitrovibrio acetiphilus* N2460, DSM 12809	646564527
Dictyoglomi	*Dictyoglomus thermophilum* H-6-12, ATCC 35947	643348542
Dictyoglomi	*Dictyoglomus turgidum* DSM 6724	643348543
Elusimicrobia	*Candidatus* Endomicrobium sp. Rs-D17	642555172
Elusimicrobia	*Elusimicrobium minutum* Pei191	642555127
Firmicutes	*Acetohalobium arabaticum* Z-7288, DSM 5501	648028002
Firmicutes	*Clostridium thermocellum* ATCC 27405	640069309
Firmicutes	*Coprothermobacter proteolyticus* DSM 5265	643348530
Firmicutes	*Halothermothrix orenii* H 168	643348557
Firmicutes	*Mesoplasma florum* L1	637000158
Firmicutes	*Moorella thermoacetica* ATCC 39073	637000167
Firmicutes	*Mycoplasma mobile* 163K	637000180
Firmicutes	*Pelotomaculum thermopropionicum* SI	640427128
Firmicutes	*Syntrophothermus lipocalidus* DSM 12680	646564577
Fusobacteria	*Fusobacterium gonidiaformans* ATCC 25563	645951804
Fusobacteria	*Fusobacterium mortiferum* ATCC 9817	646206254
Fusobacteria	*Fusobacterium nucleatum nucleatum* ATCC 23726	647000254
Fusobacteria	*Fusobacterium periodonticum* ATCC 33693	645951848
Fusobacteria	*Fusobacterium* sp. 11_3_2	651324032
Fusobacteria	*Fusobacterium ulcerans* ATCC 49185	645951859
Fusobacteria	*Fusobacterium varium* ATCC 27725	646206275
Fusobacteria	*Leptotrichia buccalis* C-1013-b, DSM 1135	644736384
Fusobacteria	*Leptotrichia goodfellowii* F0264	647000268
Fusobacteria	*Leptotrichia hofstadii* F0254	645951860
Fusobacteria	*Sebaldella termitidis* ATCC 33386	646311952
Fusobacteria	*Streptobacillus moniliformis* 9901, DSM 12112	646311956

Gemmatimonadetes	*Gemmatimonas aurantiaca* T-27T	643692024
Lentisphaerae	*Lentisphaera araneosa* HTCC2155	640963040
Nitrospirae	*Thermodesulfovibrio yellowstonii* DSM 11347	643348581
Planctomycetes	*Blastopirellula marina* SH 106T, DSM 3645	638341020
Planctomycetes	*Candidatus* Kuenenia stuttgartiensis	642555116
Planctomycetes	*Gemmata obscuriglobus* UQM 2246	641736268
Planctomycetes	*Isosphaera pallida* IS1B, ATCC 43644	649633058
Planctomycetes	*Pirellula staleyi* DSM 6068	646311948
Planctomycetes	*Planctomyces brasiliensis* IFAM 1448, DSM 5305	649633083
Planctomycetes	*Planctomyces limnophilus* Mu 290, DSM 3776	646564559
Planctomycetes	*Planctomyces maris* DSM 8797	640963032
Planctomycetes	*Rhodopirellula baltica* SH 1	637000236
Proteobacteria	*Alcanivorax borkumensis* SK2	637000004
Proteobacteria	*Anaeromyxobacter dehalogenans* 2CP-1	643348507
Proteobacteria	*Anaplasma phagocytophilum* HZ	637000009
Proteobacteria	*Arcobacter nitrofigilis* DSM 7299	646564506
Proteobacteria	*Azoarcus* sp. BH72	639633007
Proteobacteria	*Bartonella bacilliformis* KC583	639633009
Proteobacteria	*Bordetella bronchiseptica* RB50	637000032
Proteobacteria	*Brucella melitensis* ATCC 23457	643692012
Proteobacteria	*Burkholderia cenocepacia* AU 1054	637000046
Proteobacteria	*Caminibacter mediatlanticus* TB-2	640963039
Proteobacteria	*Campylobacter concisus* 13826	640753009
Proteobacteria	*Campylobacter fetus fetus* 82-40	639633016
Proteobacteria	*Campylobacter lari* RM2100	643692014
Proteobacteria	*Candidatus* Puniceispirillum marinum IMCC1322	646564516
Proteobacteria	*Cellvibrio japonicus* Ueda107	642555119
Proteobacteria	*Chromohalobacter salexigens* 1H11, DSM 3043	637000075
Proteobacteria	*Comamonas testosteroni* CNB-1	646564523
Proteobacteria	*Cupriavidus taiwanensis* LMG 19424	644736347
Proteobacteria	*Dechloromonas aromatica* RCB	637000088
Proteobacteria	*Desulfarculus baarsii* 2st14, DSM 2075	648028023
Proteobacteria	*Desulfobacterium autotrophicum* HRM2, DSM 3382	643692021
Proteobacteria	*Desulfohalobium retbaense* HR100, DSM 5692	644736349
Proteobacteria	*Desulfomicrobium baculatum* X, DSM 4028	644736350
Proteobacteria	*Desulfonatronospira thiodismutans* ASO3-1	643886196
Proteobacteria	*Desulfovibrio magneticus* RS-1	644736352
Proteobacteria	*Desulfovibrio vulgaris* Miyazaki F	643348539
Proteobacteria	*Desulfurivibrio alkaliphilus* AHT2	646564528
Proteobacteria	*Dichelobacter nodosus* VCS1703A	640427112
Proteobacteria	*Dinoroseobacter shibae* DFL-12, DSM 16493	641228491
Proteobacteria	*Erwinia amylovora* CFBP1430	646564531
Proteobacteria	*Erythrobacter litoralis* HTCC2594	637000103

Proteobacteria	*Escherichia coli* 55989	643348544
Proteobacteria	*Francisella philomiragia philomiragia* ATCC 25017	641522628
Proteobacteria	*Geobacter sulfurreducens* KN400	648231707
Proteobacteria	*Geobacter uraniireducens* Rf4	640427115
Proteobacteria	*Haliangium ochraceum* SMP-2, DSM 14365	646311933
Proteobacteria	*Helicobacter felis* CS1, ATCC 49179	649633054
Proteobacteria	*Helicobacter mustelae* ATCC 43772	646564537
Proteobacteria	*Helicobacter pullorum* MIT 98-5489	643886218
Proteobacteria	*Herbaspirillum seropedicae* SmR1	648028033
Proteobacteria	*Hyphomicrobium denitrificans* ATCC 51888	648028034
Proteobacteria	*Hyphomonas neptunium* ATCC 15444	637000135
Proteobacteria	*Kangiella koreensis* SW-125, DSM 16069	644736377
Proteobacteria	*Laribacter hongkongensis* HLHK9	643692026
Proteobacteria	*Lawsonia intracellularis* PHE/MN1-00	637000145
Proteobacteria	*Legionella longbeachae* NSW150	648028038
Proteobacteria	*Magnetococcus* sp. MC-1	639633036
Proteobacteria	*Magnetospirillum magneticum* AMB-1	637000155
Proteobacteria	*Mariprofundus ferrooxydans* PV-1	639857004
Proteobacteria	*Methylobacterium populi* BJ001	642555139
Proteobacteria	*Myxococcus fulvus* HW-1	650716065
Proteobacteria	*Nautilia profundicola* Am-H	643692029
Proteobacteria	*Neisseria lactamica* 020-06	649633075
Proteobacteria	*Neorickettsia risticii* Illinois	644736395
Proteobacteria	*Nitratifractor salsuginis* E9I37-1, DSM 16511	649633076
Proteobacteria	*Nitratiruptor* sp. SB155-2	640753037
Proteobacteria	*Nitrobacter winogradskyi* Nb-255	637000193
Proteobacteria	*Nitrosococcus watsoni* C-113	648028046
Proteobacteria	*Nitrosomonas europaea* ATCC 19718	637000195
Proteobacteria	*Nitrosospira multiformis* ATCC 25196	637000197
Proteobacteria	*Paracoccus denitrificans* PD1222	639633048
Proteobacteria	*Parvibaculum lavamentivorans* DS-1	640753040
Proteobacteria	*Parvularcula bermudensis* HTCC2503	648028050
Proteobacteria	*Pasteurella multocida multocida* Pm70	637000203
Proteobacteria	*Pelobacter carbinolicus* DSM 2380	637000204
Proteobacteria	*Pelobacter propionicus* DSM 2379	639633050
Proteobacteria	*Phenylobacterium zucineum* HLK1	642555147
Proteobacteria	*Polaromonas* sp. JS666	637000208
Proteobacteria	*Pseudomonas putida* BIRD-1	650377963
Proteobacteria	*Psychrobacter* sp. PRwf-1	640427134
Proteobacteria	*Pusillimonas* sp. T7-7	650716078
Proteobacteria	*Ralstonia pickettii* 12D	644736400
Proteobacteria	*Rhizobium rhizogenes* K84	643348504
Proteobacteria	*Rickettsia bellii* OSU 85-389	640753044
Proteobacteria	*Shewanella amazonensis* SB2B	639633057
Proteobacteria	*Starkeya novella* DSM 506	648028054
Proteobacteria	*Sulfurimonas autotrophica* OK10, DSM 16294	648028058
Proteobacteria	*Sulfurospirillum deleyianum* 5175, DSM 6946	646311960

Proteobacteria	*Syntrophobacter fumaroxidans* MPOB	639633063
Proteobacteria	*Syntrophus aciditrophicus* SB	637000317
Proteobacteria	*Thiobacillus denitrificans* ATCC 25259	637000324
Proteobacteria	*Thiomonas intermedia* K12	646564585
Proteobacteria	*Tolumonas auensis* TA 4, DSM 9187	643692052
Proteobacteria	*Vibrio furnissii* 2510/74, NCTC 11218	650377984
Proteobacteria	Wolbachia endosymbiont of Culex quinquefasciatus Pel	642555168
Proteobacteria	*Yersinia pseudotuberculosis* IP 31758	640753060
Spirochaetes	*Borrelia hermsii* DAH	642555108
Spirochaetes	*Borrelia spielmanii* A14S	642791612
Spirochaetes	*Borrelia valaisiana* VS116	641736181
Spirochaetes	*Brachyspira murdochii* 56-150, DSM 12563	646564514
Spirochaetes	*Leptospira borgpetersenii* sv Hardjo-bovis JB197	639633032
Spirochaetes	*Spirochaeta smaragdinae* SEBR 4228, DSM 11293	648028052
Spirochaetes	*Spirochaeta* sp. Buddy	650377973
Spirochaetes	*Treponema azotonutricium* ZAS-9	650716099
Spirochaetes	*Treponema brennaborense* DSM 12168	650716100
Spirochaetes	*Treponema phagedenis* F0421	649990026
Spirochaetes	*Treponema vincentii* ATCC 35580	645951869
Synergistetes	*Aminobacterium colombiense* ALA-1, DSM 12261	646564503
Synergistetes	*Jonquetella anthropi* E3_33 E1	645951855
Synergistetes	*Thermanaerovibrio acidaminovorans* Su883, DSM 6589	646311961
Thermi	*Deinococcus deserti* VCD115	643692020
Thermi	*Deinococcus geothermalis* DSM 11300	641228488
Thermi	*Deinococcus maricopensis* LB-34, DSM 21211	649633034
Thermi	*Deinococcus proteolyticus* MRP, DSM 20540	649633035
Thermi	*Deinococcus radiodurans* USUHS (R1)	637000092
Thermi	*Meiothermus ruber* 21, DSM 1279	646564545
Thermi	*Meiothermus silvanus* VI-R2, DSM 9946	646564546
Thermi	*Oceanithermus profundus* 506, DSM 14977	649633077
Thermi	*Thermus scotoductus* SA-01, ATCC 700910	649633105
Thermi	*Thermus thermophilus* HB27	637000322
Thermi	*Truepera radiovictrix* RQ-24, DSM 17093	646564586
Thermotogae	*Fervidobacterium nodosum* Rt17-B1	640753026
Thermotogae	*Kosmotoga olearia* TBF 19.5.1	644736379
Thermotogae	*Marinitoga piezophila* KA3	647533182
Thermotogae	*Mesotoga prima* MesG1.Ag.4.2	648276752
Thermotogae	*Petrotoga mobilis* SJ95	641228500
Thermotogae	*Thermosipho africanus* TCF52B	643348583
Thermotogae	*Thermosipho melanesiensis* BI429	640753057
Thermotogae	*Thermotoga lettingae* TMO	641228511
Thermotogae	*Thermotoga naphthophila* RKU-10	646311964
Thermotogae	*Thermotoga neapolitana* DSM 4359	643348584
Thermotogae	*Thermotoga petrophila* RKU-1	640427150

Verrucomicrobia	*Akkermansia muciniphila* ATCC BAA-835	642555104
Verrucomicrobia	*Chthoniobacter flavus* Ellin428	642791618
Verrucomicrobia	*Methylacidiphilum infernorum* V4	642555138
Verrucomicrobia	*Opitutus terrae* PB90-1	641522643
Verrucomicrobia	*Verrucomicrobiales* sp. DG1235	647533243
Verrucomicrobia	*Verrucomicrobium spinosum* DSM 4136	641736179
Firmicutes	*Anaerococcus prevotii* PC 1, DSM 20548	644736326
Firmicutes	*Lactobacillus gasseri* ATCC 33323	639633030
Firmicutes	*Listeria welshimeri* sv 6b, SLCC5334	639633035
Firmicutes	*Staphylococcus carnosus carnosus* TM300	643692037
Proteobacteria	*Acidithiobacillus caldus* SM-1	650716003
Proteobacteria	*Acidithiobacillus ferrooxidans* ATCC 23270	643348501
Proteobacteria	*Bacteriovorax marinus* SJ	650377909
Proteobacteria	*Bdellovibrio bacteriovorus* HD100	637000030
Proteobacteria	*Buchnera aphidicola* (Cinara tujafilina)	650716012
Proteobacteria	*Candidatus* Blochmannia floridanus	637000056
Proteobacteria	Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	637000338
Spirochaetes	*Candidatus* Cloacamonas acidaminovorans	642555115
Aquificae	*Sulfurihydrogenibium yellowstonense* SS-5	645058708
Caldiserica	*Caldisericum exile* AZM16c01	2513237181
Caldithrix	*Caldithrix abyssi* DSM 13497	2513237181
Chlorobi	*Chlorobium ferrooxidans* DSM 13031	638341060
Chrysiogenetes	*Chrysiogenes arsenatis* DSM 11915	2005520001
Crenarchaeota	*Acidilobus saccharovorans* 345-15	648028003
Crenarchaeota	*Desulfurococcus kamchatkensis* 1221n	643348540
Crenarchaeota	*Desulfurococcus mucosus* 07/1, DSM 2162	649633040
Crenarchaeota	*Hyperthermus butylicus* DSM 5456	640069314
Crenarchaeota	*Ignicoccus hospitalis* KIN4/I, DSM 18386	640753029
Crenarchaeota	*Ignisphaera aggregans* AQ1.S1, DSM 17230	648028035
Crenarchaeota	*Metallosphaera sedula* DSM 5348	640427120
Crenarchaeota	*Pyrobaculum calidifontis* JCM 11548	640069326
Crenarchaeota	*Staphylothermus marinus* F1, DSM 3639	640069332
Crenarchaeota	*Sulfolobus islandicus* HVE10/4	650377981
Crenarchaeota	*Sulfolobus tokodaii* 7, JCM 10545	638154519
Crenarchaeota	*Thermofilum pendens* Hrk 5	639633064
Crenarchaeota	*Thermoproteus uzoniensis* 768-20	650716098
Crenarchaeota	*Thermosphaera aggregans* M11TL, DSM 11486	646564583
Crenarchaeota	*Vulcanisaeta moutnovskia* 768-28	650377985
Cyanobacteria	*Arthrospira platensis* C1	2507262036
Cyanobacteria	*Cyanobacterium* sp JSC-1	2502171143
Cyanobacteria	*Fischerella* sp. JSC-11	2505679024
Cyanobacteria	*Oscillatoriales* sp. JSC-12	2510065010
Cyanobacteria	*Synechococcus elongatus* PCC 7942	2514885031
Cyanobacteria	*Synechococcus* sp. PE A4 65AY6A5	2512875021
Cyanobacteria	*Synechococcus* sp. WH 8016	2507262052
Euryarchaeota	*Aciduliprofundum boonei* T469	646564501
Euryarchaeota	*Archaeoglobus fulgidus* VC-16, DSM 4304	638154502
Euryarchaeota	*Archaeoglobus profundus* Av18, DSM 5631	646311906

Euryarchaeota	*Candidatus* Methanoregula boonei 6A8	640753014
Euryarchaeota	*Ferroglobus placidus* AEDII12DO, DSM 10642	646564534
Euryarchaeota	*Halalkalicoccus jeotgali* B3, DSM 18796	648028029
Euryarchaeota	*Haloarcula marismortui* ATCC 43049	638154503
Euryarchaeota	*Halobacterium* sp. NRC-1	638154504
Euryarchaeota	*Haloferax volcanii* DS2, ATCC 29605	646564536
Euryarchaeota	*Halogeometricum borinquense* PR3, DSM 11551	649633053
Euryarchaeota	*Halomicrobium mukohataei* arg-2, DSM 12286	644736372
Euryarchaeota	*Haloquadratum walsbyi* C23	651053028
Euryarchaeota	*Halorhabdus utahensis* AX-2, DSM 12940	644736373
Euryarchaeota	*Halorubrum lacusprofundi* ATCC 49239	643692025
Euryarchaeota	*Haloterrigena turkmenica* VKM B-1734, DSM 5511	646311934
Euryarchaeota	*Methanobacterium* sp. AL-21	650716052
Euryarchaeota	*Methanobrevibacter ruminantium* M1	646311943
Euryarchaeota	*Methanobrevibacter smithii* PS, ATCC 35061	640427121
Euryarchaeota	*Methanocaldococcus fervens* AG86	644736385
Euryarchaeota	*Methanocaldococcus infernus* ME	646564547
Euryarchaeota	*Methanocaldococcus jannaschii* DSM 2661	638154505
Euryarchaeota	*Methanocaldococcus* sp. FS406-22	646564548
Euryarchaeota	*Methanocaldococcus vulcanius* M7, DSM 12094	646311944
Euryarchaeota	*Methanocella paludicola* SANAE	646311945
Euryarchaeota	*Methanocella* sp. RC-I	640427153
Euryarchaeota	*Methanococcoides burtonii* DSM 6242	637000161
Euryarchaeota	*Methanococcus aeolicus* Nankai-3	640753034
Euryarchaeota	*Methanococcus maripaludis* C5	640069316
Euryarchaeota	*Methanococcus vannielii* SB	640753036
Euryarchaeota	*Methanococcus voltae* A3	646564549
Euryarchaeota	*Methanocorpusculum labreanum* Z	640069317
Euryarchaeota	*Methanoculleus marisnigri* JR1, DSM 1498	640069318
Euryarchaeota	*Methanohalobium evestigatum* Z-7303, DSM 3721	648028039
Euryarchaeota	*Methanohalophilus mahii* SLP, DSM 5219	646564550
Euryarchaeota	*Methanoplanus petrolearius* SEBR 4847, DSM 11571	648028040
Euryarchaeota	*Methanopyrus kandleri* AV19	638154507
Euryarchaeota	*Methanosaeta concilii* GP6	650716054
Euryarchaeota	*Methanosaeta thermophila* PT	639633038
Euryarchaeota	*Methanosarcina acetivorans* C2A	638154508
Euryarchaeota	*Methanosarcina barkeri* Fusaro, DSM 804	637000162
Euryarchaeota	*Methanosarcina mazei* Go1, DSM 3647	638154509
Euryarchaeota	*Methanosphaera stadtmanae* DSM 3091	637000163
Euryarchaeota	*Methanosphaerula palustris* E1-9c, DSM 19958	643348525
Euryarchaeota	*Methanospirillum hungatei* JF-1	637000164
Euryarchaeota	*Methanothermobacter marburgensis* Marburg	648028041

	DSM 2133	
Euryarchaeota	*Methanothermobacter thermoautotrophicus* Delta H	638154510
Euryarchaeota	*Methanothermococcus okinawensis* IH1	650716055
Euryarchaeota	*Methanothermus fervidus* V24S, DSM 2088	649633067
Euryarchaeota	*Methanotorris igneus* Kol 5	650716056
Euryarchaeota	*Natrialba magadii* ATCC 43099	646564555
Euryarchaeota	*Natronomonas pharaonis* Gabara, DSM 2160	637000187
Euryarchaeota	*Picrophilus torridus* DSM 9790	638154512
Euryarchaeota	*Pyrococcus abyssi* GE5	638154514
Euryarchaeota	*Pyrococcus furiosus* DSM 3638	638154515
Euryarchaeota	*Pyrococcus horikoshii* OT3	638154516
Euryarchaeota	*Pyrococcus* sp. NA2	650716079
Euryarchaeota	*Pyrococcus yayanosii* CH1	650716080
Euryarchaeota	*Thermococcus barophilus* MP	650716096
Euryarchaeota	*Thermococcus gammatolerans* EJ3	644736411
Euryarchaeota	*Thermococcus kodakarensis* KOD1	638154520
Euryarchaeota	*Thermococcus onnurineus* NA1	643348580
Euryarchaeota	*Thermococcus sibiricus* MM 739	644736412
Euryarchaeota	*Thermococcus* sp. 4557	650716097
Euryarchaeota	*Thermoplasma acidophilum* DSM 1728	638154521
Euryarchaeota	*Thermoplasma volcanium* GSS1	638154522
Fibrobacteres	*Fibrobacter succinogenes succinogenes* S85	650377942
Firmicutes	*Eubacterium cylindroides* T2-87	650377935
Firmicutes	*Thermodesulfobium narugense* Na82, DSM 14796	2504756006
Fusobacteria	*Fusobacterium necrophorum funduliforme* 1_1_36S	2513237330
Fusobacteria	*Fusobacterium* sp. oral taxon 370 str. F0437	2513237336
Fusobacteria	*Ilyobacter polytropus* CuHBu1, DSM 2926	649633056
Fusobacteria	*Leptotrichia goodfellowii* LB 57, DSM 19756	2506520045
Fusobacteria	*Leptotrichia shahii* DSM 19757	2515154071
Fusobacteria	*Leptotrichia wadei* DSM 19758	2515154120
Korarchaeota	*Candidatus* Korarchaeum cryptofilum OPF8	641522611
Nanoarchaeota	*Nanoarchaeum equitans* Kin4-M	638154511
Nitrospirae	*Candidatus* Nitrospira defluvii	649633030
Nitrospirae	*Leptospirillum ferrooxidans* C2-3	2540341086
OP10	*Chthonomonas calidirosea* T49	2503242004
OP9	*Caldatribacterium* OP9-cSCG	APKF00000000
Saccharibacteria	*Candidatus* Saccharimonas aalborgensis	CP005957.1
Thaumarchaeota	*Cenarchaeum symbiosum* A	641522613
Thaumarchaeota	*Nitrosopumilus maritimus* SCM1	641228499
Verrucomicrobia	*Opitutaceae* sp. TAV2	640963002
WWE1	*Candidatus* Cloacamonas acidaminovorans	642555115

Table S2.5. Genes belong to pathways or assemblages from the Melainabacteria representatives from this study
Listed are the IMG gene IDs for each of the Melainabacteria that are deposited in IMG.

		Symbol	*Candidatus* Gastroanaerophila phascolarctos	Zag_1	Zag_111	MH_37	*Candidatus* Obscuribacter phosphatis	*Candidatus* Caenarcanophila bioreactus
EMP pathway	**E.C number**							
Glucokinase	2.7.1.2	GK	2523622275	2523618351	2534657128 2534657723	2522811268	2541282499 2541283255	-
Glucose-6-phosphate isomerase	5.3.1.9	GPI	2523621784	2523618311	2534656686	2522811348	-	2554235349
Phosphofructokinase	2.7.1.11	PFK	-	-	2534656845	-	2541281805	2554235646
Fructose-bisphosphate aldolase	4.1.2.13	fbaA	2523623187	2523617884	2534656936	2522812428	2541285301	2554235409
Triosephosphate isomerase	5.3.1.1	TPI	2523623082	2523617328	2534658147	2522812616	2541284414	2554236361
Glyceraldehyde-3-phosphate dehydrogenase	1.2.1.12	GAP	2523622612	2523617047	2534657629	-	2541285214	2554236616
Phosphoglycerate kinase	2.7.2.3	PGK	2523621953	2523618679	2534656945	2522812014	2541285215	2554235911
Phosphoglycerate mutase	5.4.2.1	PGM	2523622862	2523618262	2534656858	2522813130	2541282187	2554235338
Enolase	4.2.1.11	ENO	2523623115	2523618038	2534656767	2522811435	2541282164	-
Pyruvate kinase	2.7.1.40	PK	2523622478	2523618904	2534656599	2522811367	2541285471	2554235818
Fermentation								
Acetaldehyde dehydrogenase/alcohol dehydrogenase	1.2.1.10/1.1.1.1	ALDH/ADH	2523622361	2523617074	2534658922	2522812406	2541282950	2554235808
Alcohol dehydrogenase, class IV	1.1.1.1	ADH	2523621842	2523617291	2534657773	2522812145	2541282501	2554235925
Lactate dehydrogenase	1.1.1.28	LDH	2523621888	2523618732	2534657873	2522811332	2541281883	2554236386
Pyruvate formate lyase	2.3.1.54	PFL	2523622898	-	2534656682	-	-	-
Pentose phosphate pathway								
Glucose-6-phosphate dehydrogenase	1.1.1.49	G6PD	-	-	-	-	2541282512	-

6-phosphogluconolactonase	3.1.1.31	PGLS	-	-	-	-	2541282514	-
6-phosphogluconate dehydrogenase	1.1.1.44	PGD	-	-	-	-	2541282515	-
Ribulose-5-phosphate 3-Epimerase	5.1.3.1	RPE	2523623171	2523619049	2534656104 2534657714	2522811657	2541282203	2554235427
Ribose-5-phosphate Isomerase	5.3.1.6	RPI	2523621712	2523619138	2534657230	2522811997	2541282259	2554235504
Transketolase	2.2.1.1	TKT	2523622052 2523621794	2523618017 2523618410	2534656618 2534657589	2522812113 2522812117 2522812264 2522812114 2522812118 2522813495	2541283291	2554236287 2554235334
Transaldolase	2.2.1.2	TAL	-	-	-	-	2541281904	2554236185
TCA cycle								
Pyruvate dehydrogenase	1.2.4.1	PDK	-	-	-	-	2541283534 2541283535	-
Citrate synthase	2.3.3.1	CS	-	-	-	-	2541282238	-
Aconitase	4.2.1.3	ACO	-	-	-	-	2541284223	-
Isocitrate dehydrogenase	1.1.1.41	IDH	-	2523618070	2534657368	2522811358		2554236132
Isocitrate dehydrogenase	1.1.1.42	IDH	-	-	-	-	2541284187 2541284595	-
2-oxoacid:ferredoxin oxidoreductase	1.2.7.3	OFOR	2523622373 2523622374 2523622375 2523622658 2523622659 2523622660	2523617943 2523617944 2523617945	2534657734 2534657735 2534657736	2522812645 2522812646 2522812647	2541281896 2541281897 2541282395 2541282396	-
Succinyl-CoA synthetase	6.2.1.5	SCS	-	-	-	-	2541282222 2541282223	-
Succinate dehydrogenase	1.3.5.1	SDH	-	-	-	-	2541284229 2541284230	-
Fumarase	4.2.1.2	FH	-	-	-	-	2541282719	-
Malate dehydrogenase	1.1.1.37	MDH	-	-	-	-	2541284440	-
Phosphoenolpyruvate	4.1.1.32	PEPCK	-	-	-	-	2541283468	-

carboxykinase								
Electron transport chain								
Predicted nucleoside-diphosphate-sugar epimerases	1.6.99.3	-	-	-	-	-	2541282363 2541284393	-
NADH dehydrogenase subunit A	1.6.5.3	nuoA	-	-	-	-	2541285507	-
NADH dehydrogenase subunit B	1.6.5.3	nuoB	-	-	-	-	2541285508	-
NADH:ubiquinone oxidoreductase 27 kD subunit	1.6.5.3	nuoC	-	-	-	-	2541285509	-
NADH:ubiquinone oxidoreductase 49 kD subunit 7	1.6.5.3	nuoD	-	-	-	-	2541285510	-
NADH:ubiquinone oxidoreductase 24 kD subunit	1.6.5.3	nuoE	-	2523617688	2534659396	2522812754	-	2554235204
NAD(P)-dependent iron-only hydrogenase diaphorase component flavavoprotein	1.6.5.3	nuoF	-	2523617687	2534659397	2522812755	-	2554235203
NAD(P)-dependent iron-only hydrogenase catalytic subunit	1.6.5.3	nuoG	-	2523617686	2534659398	2522812756	-	2554235202
NADH dehydrogenase subunit H	1.6.5.3	nuoH	-	-	-	-	2541285511	-
NADH-quinone oxidoreductase, chain I	1.6.5.3	nuoI	-	-	-	-	2541285512	-
NADH dehydrogenase subunit L	1.6.5.3	nuoL	-	-	-	-	2541281894	-
Proton-translocating NADH-quinone oxidoreductase, chain M	-	nuoM	-	-	-	-	2541281893	-
NADH dehydrogenase subunit N	1.6.5.3	nuoN	-	-	-	-	2541283929	-
NADH:ubiquinone oxidoreductase subunit 5	1.6.5.3	ndhF	-	-	-	-	2541284457	-

(chain L)/Multisubunit Na+/H+ antiporter, MnhA subunit								
NADH:ubiquinone oxidoreductase subunit 6 (chain J)	1.6.5.3	ndhG	-	-	-	-	2541285513	-
Succinate dehydrogenase subunit A	1.3.5.1/1.3.99.1	sdhA	-	-	-	-	2541284229	-
Succinate dehydrogenase subunit B	1.3.5.1/1.3.99.1	sdhB	-	-	-	-	2541284230	-
Cbb3-type cytochrome oxidase, cytochrome c subunit	-	ccoO	-	-	-	-	2541285061	-
Cbb3-type cytochrome oxidase, subunit 1	1.9.3.1	ccoN	-	-	-	-	2541285062	-
ATP synthase F1 subcomplex alpha subunit	3.6.3.1.4	atpA	2523622717	2523618879	2534659071	2522812078	2541284051	2554236353
ATP synthase F0 subcomplex A subunit	3.6.3.1.4	atpB	2523622712	2523619145	2534658915	2522812073	2541284046	2554235801
ATP synthase F1 subcomplex epsilon subunit	3.6.3.1.4	atpC	2523621799	2523618795	-	2522813083	2541283748	-
ATP synthase F1 subcomplex beta subunit	3.6.3.1.4	atpD	2523621800	2523618796	2534659931	2522813084	2541283746	2554236478
ATP synthase, F0 subunit C	3.6.3.1.4	atpE	2523622713	2523619144	2534659316	2522812074	2541284047	2554235285
F0-F1-type ATP synthase, subunit b	3.6.3.1.4	atpF	2523622714	2523619143	-	2522812075	2541284048 2541284049	-
ATP synthase F1 subcomplex gamma subunit	3.6.3.1.4	atpG	2523622718	2523618878	2534659807	2522812079	2541284052	2554235023
ATP synthase F1 subcomplex delta subunit	3.6.3.1.4	atpH	2523622716	-	-	2522812077	2541284050	-
ATP synthase F0 subcomplex C subunit	3.6.3.1.4	-	-	-	-	-	-	-
Inorganic pyrophosphatase	3.6.1.1	-	2523621727	2523617466	-	2522812545	2541285303	-

Polyphosphate kinase 1	2.7.4.1	ppk	-	-	-	-	2541282432	-
Vacuolar-type H(+) translocating pyrophosphatase	-	-	-	-	-	-	2541285109	-
Flagella assembly								
Flagella basal body P-ring formation protein FlgA		FlgA	-	-	-	-	-	-
flagellar basal-body rod protein FlgB		FlgB	2523621836	2523618479	2534656302	2522812505	-	-
flagellar basal-body rod protein FlgC		FlgC	2523621835	2523618478	2534656301	2522812504	-	-
Flagellar hook capping protein		FlgD	-	-	-	-	-	-
flagellar hook-basal body protein FlgE		FlgE	-	-	-	-	-	-
flagellar hook-basal body rod protein FlgF		FlgF	-	2523618481	2534656304	-	-	-
flagellar basal-body rod protein FlgG, Gram-negative bacteria		FlgG	2523621837 2523621838	2523618480	2534656303	2522812506	-	-
Flagellar basal body L-ring protein		FlgH	-	-	-	-	-	-
Flagellar basal-body P-ring protein		FlgI	-	-	-	-	-	-
Flagellar protein FlgJ		FlgJ	-	-	-	-	-	-
Flagellar hook-associated protein FlgK		FlgK	-	-	-	-	-	-
flagellar hook-associated protein 3		FlgL	-	-	-	-	-	-
FlgN		FlgN	-	-	-	-	-	-
Flagellar biosynthesis pathway, component FlhA		FlhA	2523622290	2523618233	2534658234	2522812376	-	2554236289
Flagellar biosynthesis pathway, component FlhB		FlhB	2523621565	2523617711	2534656770	2522811979	-	2554235321

Flagellin and related hook-associated proteins		FliC	-	-	-	-	-	-
Flagellar capping protein		FliD	-	-	-	-	-	-
flagellar hook-basal body complex protein FliE		FliE	2523622294	2523618229	-	2522812380	-	-
flagellar basal-body M-ring protein/flagellar hook-basal body protein (fliF)		FliF	-	-	-	-	-	-
Flagellar motor switch protein FliG		FliG	-	-	-	-	-	-
Flagellar biosynthesis/type III secretory pathway protein		FliH	-	-	-	-	-	-
type III secretion system ATPase, FliI/YscN (EC 3.6.3.15)		FliI	2523621472	2523618466 2523619149	2534657369	2522811726	-	2554236374
flagellar export protein FliJ		FliJ	2523621613	2523618709	-	2522811828	-	-
Flagellar hook-length control protein FliK		FliK	-	-	-	-	-	-
Flagellar basal body-associated protein		FliL	-	-	-	-	-	-
Flagellar motor switch protein FliM		FliM	-	-	-	-	-	-
Flagellar motor switch protein FliN		FliN/SpoA	-	-	-	-	-	-
Flagellar biosynthesis protein, FliO		FliO	-	-	-	-	-	-
Flagellar biosynthesis pathway, component FliP		FliP	2523622788	2523618755	2534657253	2522811903	-	2554235810
Flagellar biosynthesis pathway, component FliQ		FliQ	-	-	-	-	-	-
Flagellar biosynthesis pathway, component FliR		FliR	2523622786	2523618758	2534657251	2522812209	-	2554235468
flagellar biosynthetic protein FliS		FliS	-	-	-	-	-	-

Flagellar protiein FliT		FliT	-	-	-	-	-	-
Flagella motor component		MotA	-	-	-	-	-	-
Flagella motor protein		MotB	-	-	-	-	-	-
Flagellar motor protein		OmpA	-	-	2534656717	-	-	-
Hydrogenases								
[FeFe] hydrogenase H-cluster radical SAM maturase HydE	2.8.1.6	HydE	-	-	2534657619	2522813318	-	2554235784
iron-only hydrogenase maturation protein HydF	-	HydF	-	2523618604	2534656733	2522811229	-	2554235601
iron-only hydrogenase maturation protein HydG	-	HydG	-	2523617045	2534657108	2522813171	-	2554235630
[FeFe] hydrogenase, group B1/B3	-	-	-	-	2534657038	2522811239	-	-
Iron only hydrogenase large subunit, C-terminal domain	-	-	-	2523617266	2534657035	2522811236	-	-
NAD(P)-dependent iron-only hydrogenase diaphorase component flavoprotein	1.6.5.3	-	-	2523617687	2534657247	2522812755	-	2554235203
NAD(P)-dependent iron-only hydrogenase catalytic subunit	1.6.5.3	-	-	2523617686	2534657246	2522812756	-	2554235202
Ni,Fe-hydrogenase I small subunit	1.12.99.6	-	-	-	-	-	2541282828	-
Ni,Fe-hydrogenase I large subunit	1.12.99.6	-	-	-	-	-	2541282829	-
hydrogenase maturation protease	-	HycI	-	-	-	-	2541282830	-
Ni,Fe-hydrogenase III small subunit	-	-	2523623068	-	-	-	2541283385	-
Ni,Fe-hydrogenase III large subunit	-	-	2523623069	-	-	-	2541283386	-
Hydrogenase 4 membrane component (E)	1.-	-	2523623071	-	-	-	2541283388	-

Coenzyme F420-reducing hydrogenase, alpha subunit	-	-	-	-	-	-	2541281832	-
Coenzyme F420-reducing hydrogenase, beta subunit	-	-	-	-	2534657715	-	-	-
Coenzyme F420-reducing hydrogenase, delta subunit	-	-	-	-	-	-	2541285052	-
Coenzyme F420-reducing hydrogenase, gamma subunit	-	-	-	-	-	-	2541281831	-
Zn finger protein HypA/HybF (possibly regulating hydrogenase expression)	-	HypA/HybF	-	-	-	-	2541282821	-
Hydrogenase nickel incorporation protein HypB	-	HypB	-	-	-	-	2541282820	-
Hydrogenase maturation protein HypC	-	HypC	-	-	-	-	2541282822	-
Hydrogenase maturation protein HypD	-	HypD	-	-	-	-	2541282823 2541285008	-
Hydrogenase maturation protein, carbamoyl dehydratase HypE	-	HypE	-	-	-	-	2541282825	-
Hydrogenase maturation protein, carbamoyltransferase HypF	-	HypF	-	-	-	-	2541282826	-
Polyphosphate metabolism								
Polyphosphate kinase	-	-	-	-	-	2522813547	-	-
polyphosphate kinase 1	2.7.4.1	-	-	-	-	-	2541282432	-
polyphosphate kinase 2, PA0141 family	-	-	-	-	-	-	2541282023	-
Polyphosphate:AMP phosphotransferase	2.7.4.-	-	-	-	-	-	2541284980	-
Guanosine polyphosphate	3.1.7.2	-	2523622599	2523617618	2534656623	-	2541284760	-

pyrophosphohydrolases/synthetases								
Exopolyphosphatase	3.6.1.11	-	-	-	-	-	2541282351	-
Adenylate kinase	2.7.4.3	-	2523623169	2523616999	2534657048	2522813109	2541282076	2554235454

Table S2.6. Genomes used to produce the flagella gene tree
Listed are the IMG taxon ID, the organism name and the phylum that the organism belongs to.

IMG taxon ID	Organism	Phylum
643692001	*Acidobacterium capsulatum* ATCC 51196	Acidobacteria
649633100	*Terriglobus saanensis* SP1PR4, DSM 23119	Acidobacteria
642555107	*Bifidobacterium longum* DJO10A	Actinobacteria
2508501106	*Mycobacterium rhodesiae* NBB3	Actinobacteria
2517434006	*Brevibacterium casei* S18	Actinobacteria
646564582	*Thermocrinis albus* HI 11/12, DSM 14484	Aquificae
643692050	*Sulfurihydrogenibium azorense* Az-Fu1	Aquificae
649633104	*Thermovibrio ammonificans* HB-1, DSM 15698	Aquificae
2511231141	*Alicyclobacillus acidocaldarius acidocaldarius* Tc-4-1	Bacillus
649633013	*Bacteroides salanitronis* BL78, DSM 18170	Bacteroides
640753008	*Bacteroides vulgatus* ATCC 8482	Bacteroides
637000065	*Chlamydophila abortus* S26/3	Chlamydia
637000067	*Chlamydophila felis* Fe/C-56	Chlamydia
646564588	*Waddlia chondrophila* WSU 86-1044	Chlamydia
637000073	*Chlorobium tepidum* TLS	Chlorobi
637000072	*Chlorobium chlorochromatii* CaD3	Chlorobi
642555122	*Chlorobium phaeobacteroides* BS1	Chlorobi
2508501111	*Herpetosiphon aurantiacus* DSM 785	Chloroflexi
649989977	*Oscillochloris trichoides* DG6	Chloroflexi
649633005	*Anaerolinea thermophila* UN-1	Chloroflexi
649633038	*Desulfurispirillum indicum* S5, DSM 22839	Chrysiogenetes
637000076	*Clostridium acetobutylicum* ATCC 824	Clostridia
2503508009	*Mahella australiensis* 50-1 BON, DSM 15567	Clostridia
641522632	*Heliobacterium modesticaldum* Ice1	Clostridia
649633052	*Halanaerobium hydrogenoformans*	Clostridia
640427120	*Metallosphaera sedula* DSM 5348	Crenarchaeota
641228499	*Nitrosopumilus maritimus* SCM1	Crenarchaeota
648028062	*Vulcanisaeta distributa* DSM 14429	Crenarchaeota
2503982047	*Anabaena cylindrica* PCC 7122	Cyanobacteria
639857037	*Nodularia spumigena* CCY9414	Cyanobacteria
637000313	*Synechococcus* sp. JA-3-3Ab	Cyanobacteria
637000121	*Gloeobacter violaceus* PCC 7421	Cyanobacteria
2523533517	Zag_1	Cyanobacteria
2531839741	Zag_111	Cyanobacteria
2523533519	*Ca.* Gastranaerophilus phascolarctosicola	Cyanobacteria
2522572068	MH_37	Cyanobacteria
2541046959	Mel_A1	Cyanobacteria
2541046956	Mel_B1	Cyanobacteria
2541046940	Mel_B2	Cyanobacteria
2541046938	Mel_C1	Cyanobacteria

2531839742	*Ca.* Caenarcanum bioreactoricola	Cyanobacteria
2541046960	*Ca.* Obscuribacter phosphatis	Cyanobacteria
2541046958	ACD20	Cyanobacteria
643348543	*Dictyoglomus turgidum* DSM 6724	Dictyoglomi
643348542	*Dictyoglomus thermophilum* H-6-12, ATCC 35947	Dictyoglomi
642555127	*Elusimicrobium minutum* Pei191	Elusimicrobia
642555172	*Candidatus* Endomicrobium sp. Rs-D17	Elusimicrobia
644736373	*Halorhabdus utahensis* AX-2	Euryarcheota
648028040	*Methanoplanus petrolearius* SEBR 4847	Euryarcheota
638154522	*Thermoplasma volcanium* GSS1	Euryarcheota
650377942	*Fibrobacter succinogenes succinogenes* S85	Fibrobacter
637000117	*Fusobacterium nucleatum nucleatum* ATCC 25586	Fusobacteria
646311952	*Sebaldella termitidis* ATCC 33386	Fusobacteria
644736384	*Leptotrichia buccalis* C-1013-b, DSM 1135	Fusobacteria
643692024	*Gemmatimonas aurantiaca* T-27T	Gemmatimonadetes
640963040	*Lentisphaera araneosa* HTCC2155	Lentisphaerae
650633000	*Victivallis vadensis* ATCC BAA-548	Lentisphaerae
638154511	*Nanoarchaeum equitans* Kin4-M	Nanoarchaeota
2540341086	*Leptospirillum ferrooxidans* C2-3	Nitrospirae
637000236	*Rhodopirellula baltica* SH 1	Planctomycetes
649633083	*Planctomyces brasiliensis* IFAM 1448, DSM 5305	Planctomycetes
641736268	*Gemmata obscuriglobus* UQM 2246	Planctomycetes
643692004	*Azotobacter vinelandii* DJ, ATCC BAA-1303	Proteobacteria
649633004	*Alicycliphilus denitrificans* BC	Proteobacteria
650716078	*Pusillimonas* sp. T7-7	Proteobacteria
637000241	*Rhodospirillum rubrum* S1, ATCC 11170	Proteobacteria
646311920	*Dickeya dadantii* Ech586	Proteobacteria
644736355	*Dickeya zeae* Ech1591	Proteobacteria
637000207	*Photorhabdus luminescens laumondii* TTO1	Proteobacteria
650377903	*Acinetobacter calcoaceticus* PHEA-2	Proteobacteria
643348518	*Borrelia duttonii* Ly	Spirochaetes
2506783010	*Leptonema illini* 3055, DSM 21528	Spirochaetes
2511231215	*Treponema pallidum pertenue* Gauthier	Spirochaetes
2505119043	*Thermovirga lienii* Cas60314, DSM 17291	Synergistetes
646311961	*Thermanaerovibrio acidaminovorans* Su883, DSM 6589	Synergistetes
645951855	*Jonquetella anthropi* E3_33 E1	Synergistetes
2508501115	*Deinococcus pimensis* KR-235	Thermi
646564545	*Meiothermus ruber* 21, DSM 1279	Thermi
2515154172	*Thermus igniterrae* ATCC 700962	Thermi
2505119042	*Thermodesulfatator indicus* CIR29812, DSM 15286	Thermodesulfobacteria
640427150	*Thermotoga petrophila* RKU-1	Thermotoga
2510065086	*Mesotoga prima* MesG1Ag4.2	Thermotoga

2519899531	*Thermotoga maritima* MSB8, DSM 3109	Thermotoga
2517572100	*Opitutaceae* sp. TAV2	Verrucomicrobia
641522643	*Opitutus terrae* PB90-1	Verrucomicrobia
642791618	*Chthoniobacter flavus* Ellin428	Verrucomicrobia

Table S2.7. Table of Pfams used to differentiate cell wall types and flagella assembly, and GI numbers for photosynthesis genes and (bacterio)chlorophyll biosynthesis genes

The PFAM numbers that were used to differentiate cell wall were obtained from Albertsen *et al.,* 2013. The flagella assembly genes are outlined in Pallen and Matzke, 2006 and the the (bacterio)chlorophyll biosynthesis genes and cut-offs were obtained from Sousa *et al.,* 2012.

Pfam/GI	Pfam description
PF04413	Glycos_transf_N – (kdotransferase)
PF02614	LpxK – Tetraacyldisaccharide-1-P 4'-kinase
PF02684	LpxB – Lipid-A-disaccharide synthetase
PF03331	LpxC – UDP-3-O-acyl N-acetylglycosamine deacetylase
PF04613	LpxD – UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase, LpxD
PF02472	ExbD – Biopolymer transport protein ExbD/TolR
PF07244	Surf_Ag_VNR – Surface antigen variable number repeat
PF03739	YjgP_YjgQ – Predicted permease YjgP/YjgQ family
PF01103	Bac_surface_Ag – Surface antigen
PF00263	Secretin – Bacterial type II and III secretion system protein
PF02321	OEP – Outer membrane efflux protein
PF03968	OstA – OstA-like protein
PF00593	TonB_dep_Rec – TonB dependent receptor
PF04166	PdxA – Pyridoxal phosphate biosynthesis protein PdxA
PF06835	Lipopolysaccharide-assembly, LptC-related
PF03740	PdxJ – Pyridoxal phosphate biosynthesis protein PdxJ
PF03548	LolA – Outer membrane lipoprotein carrier protein LolA
PF04052	TolB_N – TolB amino-terminal domain
PF04453	OstA_C – Organic solvent tolerance protein
PF02645	DegV – Uncharacterised protein, DegV family COG1307
PF05103	DivIVA – DivIVA protein
PF02650	HTH_WhiA – Sporulation Regulator WhiA C terminal domain
PF10298	WhiA_N – Sporulation Regulator WhiA N terminal
PF04472	DUF552 – Protein of unknown function (DUF552)
PF04203	Sortase – Sortase family
PF03816	LytR_cpsA_psr – Cell envelope-related transcriptional attenuator domain
PF09269	DUF1967 – Domain of unknown function (DUF1967)
PF01424	R3H – R3H domain
PF01618	MotA_ExbB – MotA/TolQ/ExbB proton channel family
PF13677	MotB_plug – Membrane MotB of proton-channel complex MotA/MotB
PF03963	FlgD – Flagellar hook capping protein – N-terminal region
PF00460 PF06429	Flg_bb_rod – Flagella basal body rod protein Flg_bbr_C – Flagellar basal body rod FlgEFG protein C-terminal
PF02107	FlgH - Flagellar L-ring protein
PF02119	FlgI - Flagellar P-ring protein
PF00669 PF00700	Flagellin_N – Bacterial flagellin N-terminal helical region Flagellin_C – Bacterial flagellin C-terminal helical region
PF02465 PF07195	FliD_N - Flagellar hook-associated protein 2 N-terminus FliD_C – Flagellar hook-associated protein 2 C-terminus
PF02049	FliE – Flagellar hook-basal body complex protein FliE
PF01514 PF08345	YscJ_FliF – Secretory protein of YscJ/FliF family YscJ_FliF_C – Flagellar M-ring protein C-terminal
PF01706	FliG_C - FliG C-terminal domain

PF02108	FliH – Flagellar assembly protein FliH
PF02050	FliJ – Flagellar FliJ protein
PF02154	FliM - Flagellar motor switch protein FliM
PF01052	SpoA - Surface presentation of antigens (SPOA)
PF00813	FliP – FliP family
PF01313	Bac_export_3 – Bacterial export proteins, family 3
PF01311	Bac_export_1 – Bacterial export proteins, family 1
PF02561	FliS – Flagellar protein FliS
PF00771	FHIPEP - FHIPEP protein family
PF01312	Bac_export_2 – FlhB HrpN YscU SpaS Family
189347628 21674769 37522897 77463857	bchH
189346994 21674119 37520439 77463844	bchD
21674120 37521283 77463843	bchI
21674770 37523971 77463859	bchM
21674771 77463851	bchE
77463865	acsF
21673889	bciA
16331168	bciB
159462468	LPOR
189347668 21674961 37521938 77463855	bchN
21674960 37519784 77463856	bchB
189347666 21674959 37521939 77463858	bchL

Appendix B: Supplementary figures and tables for Chapter 3

Supplementary Figures

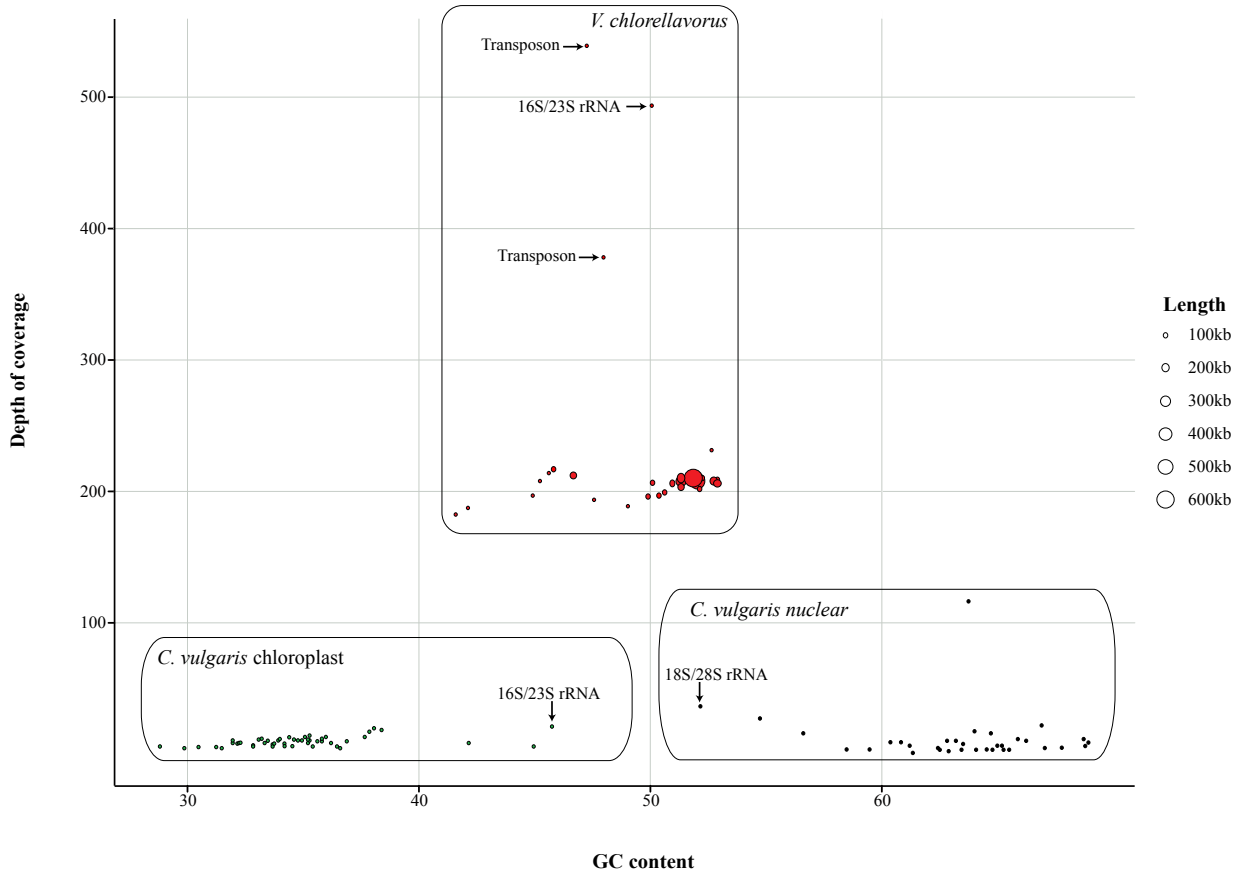


Figure S3.1. Depth of coverage against GC content for the assembled contigs. Contigs assigned to *V. chlorellavorus* are represented by red circles. *C. vulgaris* chloroplast contigs are represented by green circles and *C. vulgaris* contigs are represented by black circles. The size of the circle corresponds to the length of the contig.

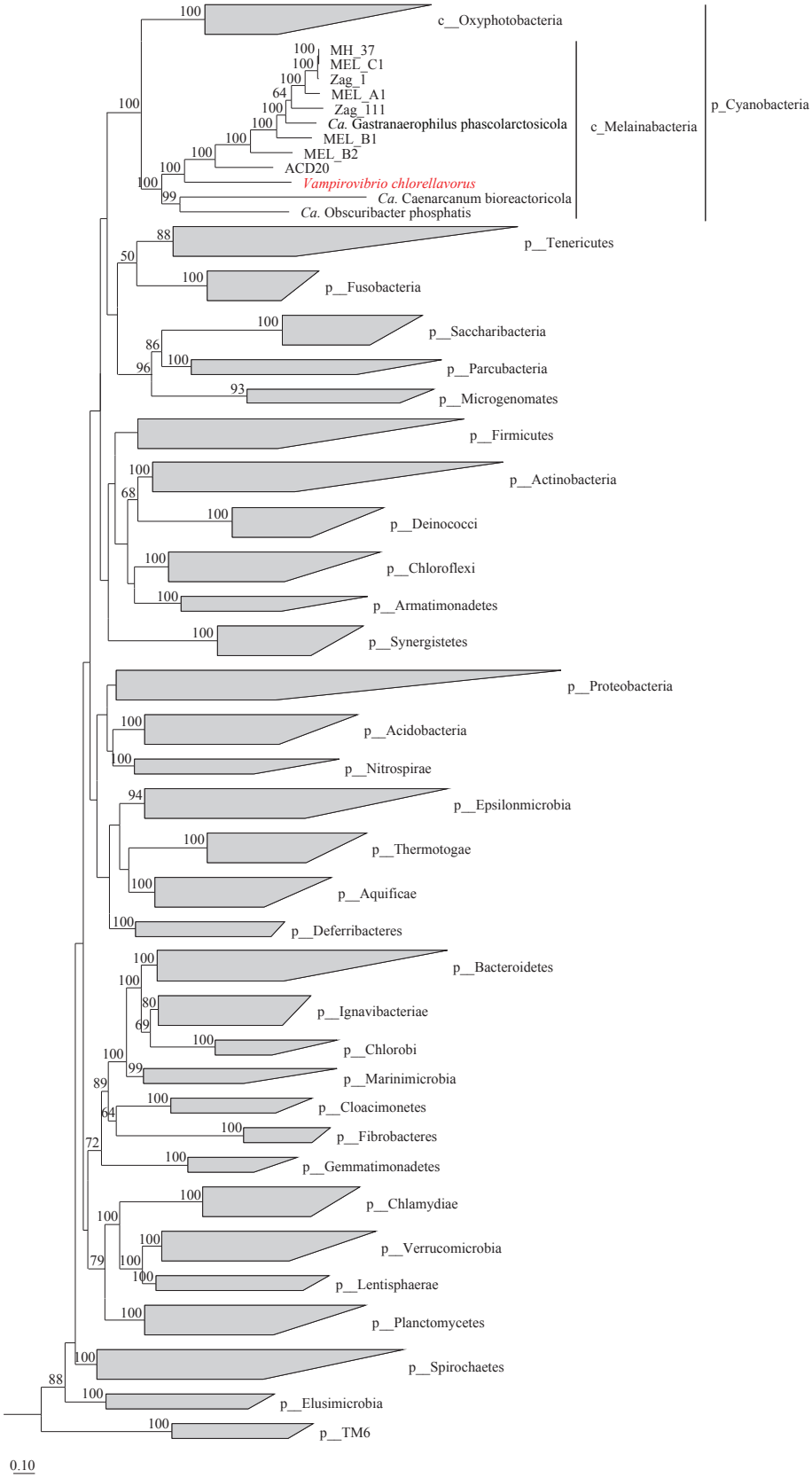


Figure S3.2. A maximum likelihood concatenated gene tree showing the Cyanobacteria and select bacterial phyla. The phylogenetic tree was inferred from the concatenation of 109 conserved marker genes (**Table S3.1**) and consists of 7,732 bacterial and 169 archaeal genomes from the IMG database (Markowitz *et al.,* 2014).

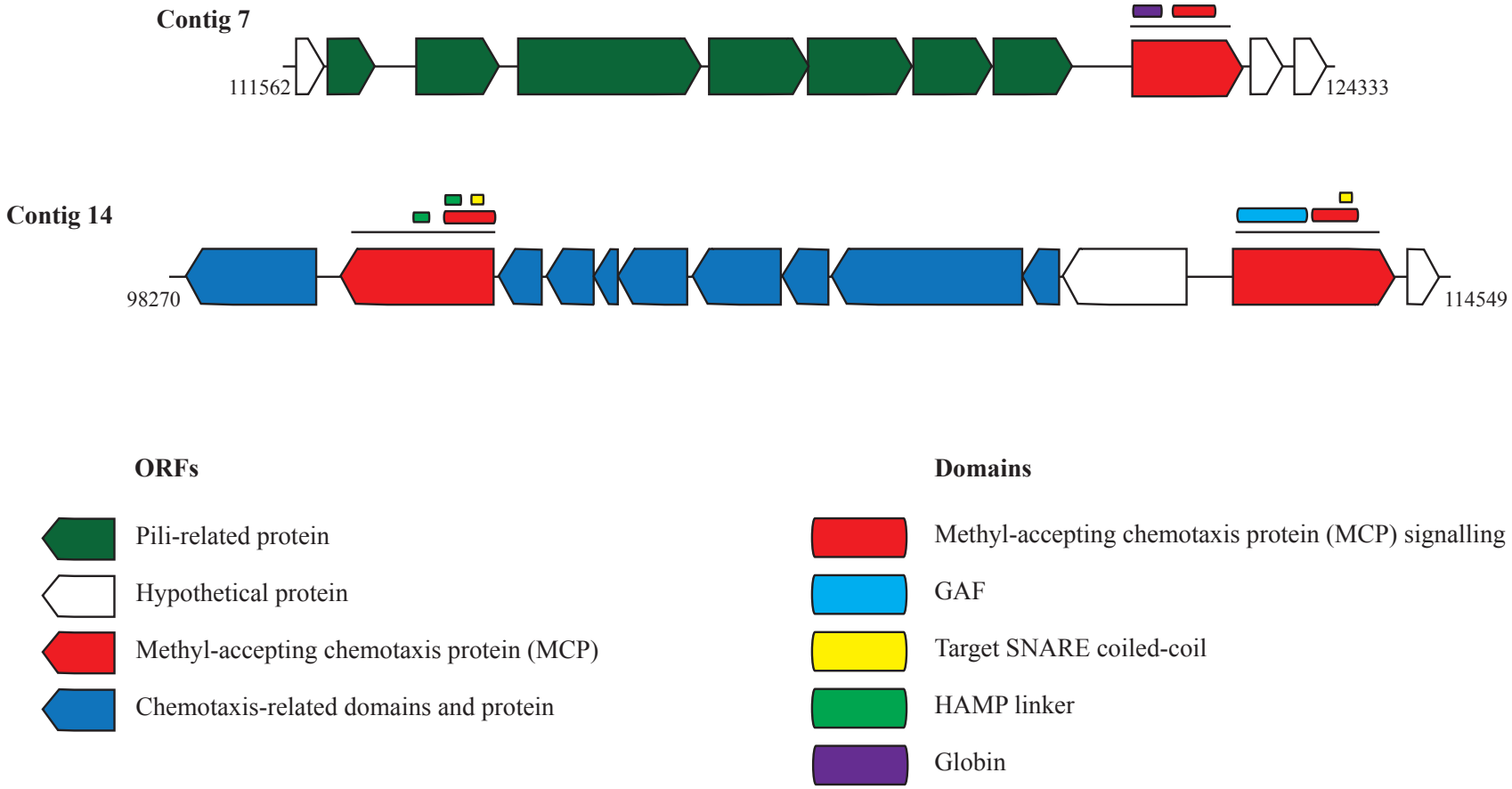


Figure S3.3. Methyl-accepting chemotaxis proteins (MCP) encoded on the *V. chlorellavorus* genome. Three MCPs are found in *V. chlorellavorus*. MCP domains were predicted with InterProScan5 (Jones *et al.,* 2014). Directional arrows represent the putative genes and orientation (positive or negative strand) and the rounded rectangles represent domains.

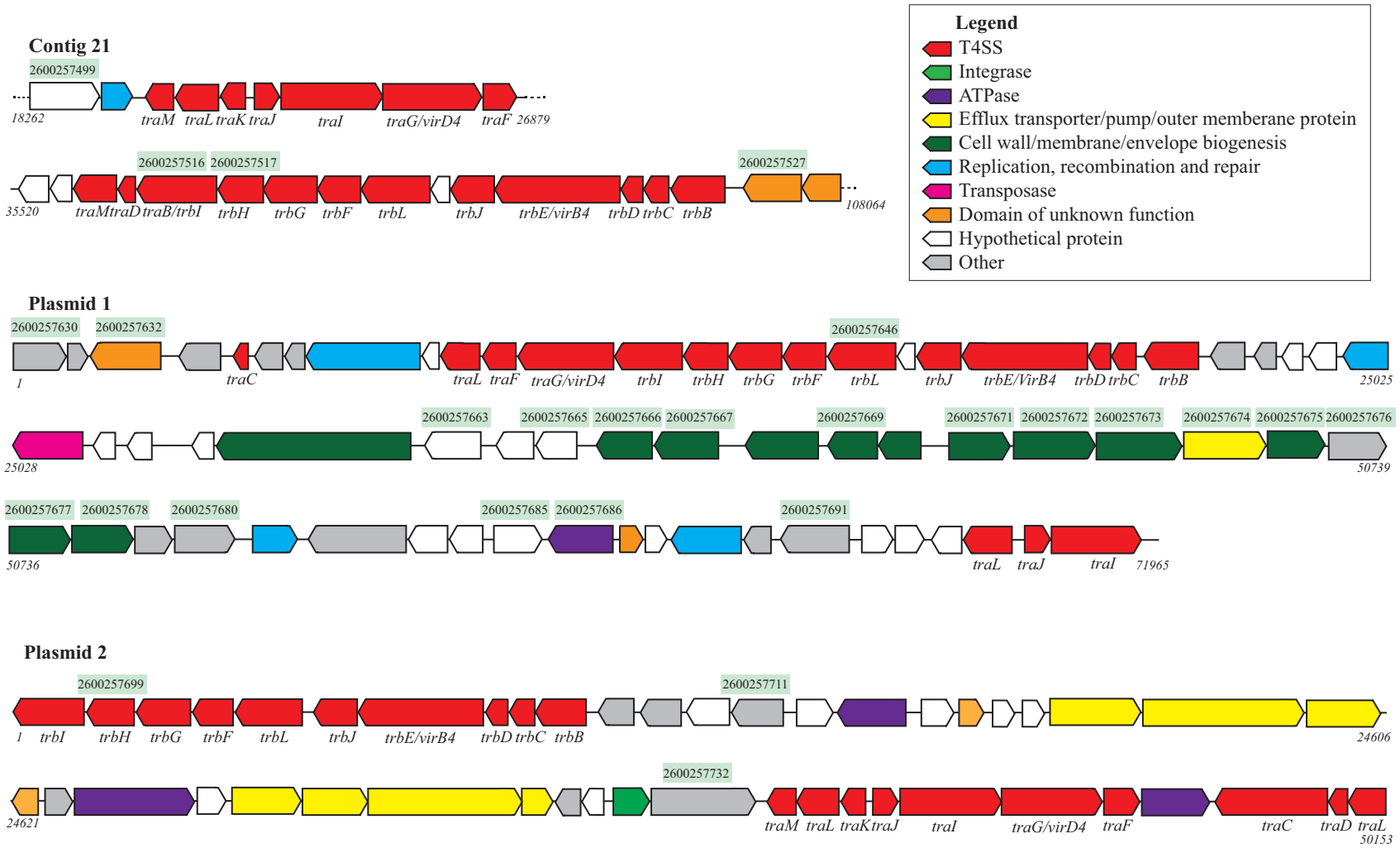


Figure S3.4. Type IV secretion system (T4SS) in the *V. chlorellavorus* genome and plasmids. The schematic diagram shows the presence of T4SS genes identified by IMG/ER on both plasmids and one contig. The arrows represent annotated genes and their direction. The numbers above the genes are the IMG/ER accession numbers and these have been identified as alien genes by PHX analysis (**Table S3.2**). T4SS genes have the predicted gene names italicised below.

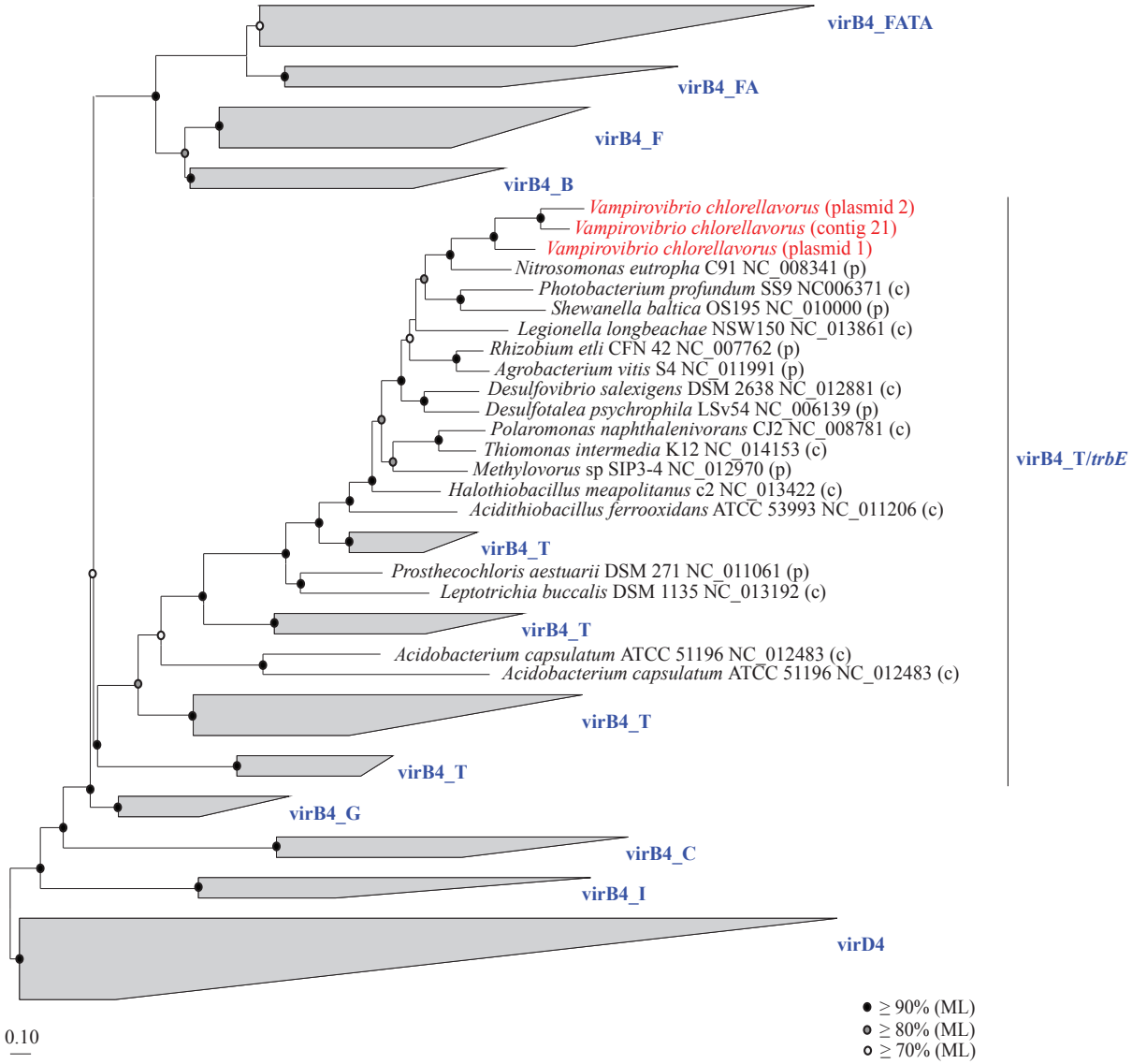


Figure S3.5. A maximum likelihood phylogenetic tree of *virB4* genes. Aligned sequences and naming conventions were obtained from Christie, 2004. virB4_T is based on the T-DNA conjugation system of *Agrobacterium tumefaciens* plasmid Ti, virB4_F is based on the plasmid F, virB4_I is based on the Incl plasmid R64 and virB4_G is based on ICEHIN1056. The other T4SS have homologues to VirB4 and include the Cyanobacteria (virB4_C), Bacteroides (virB4_B), Firmicutes (virB4_FA and virB4_FATA), Actinobacteria (virB4_FA and virB4_FATA), Tenericutes (virB4_FATA) and Archaea (virB4_FATA) (Christie, 2004). The *V. chlorellavorus* genome contains T4SS that belong to the virB4_T. virD4 is used as the outgroup. Black circles represent nodes with ≥ 90% bootstrap support, grey circles represent nodes with ≥ 80% bootstrap support and white circles represent nodes with ≥ 70% bootstrap support. (p) corresponds to *virB4* genes found on plasmids and (c) corresponds to *virB4* genes found on the chromosome.

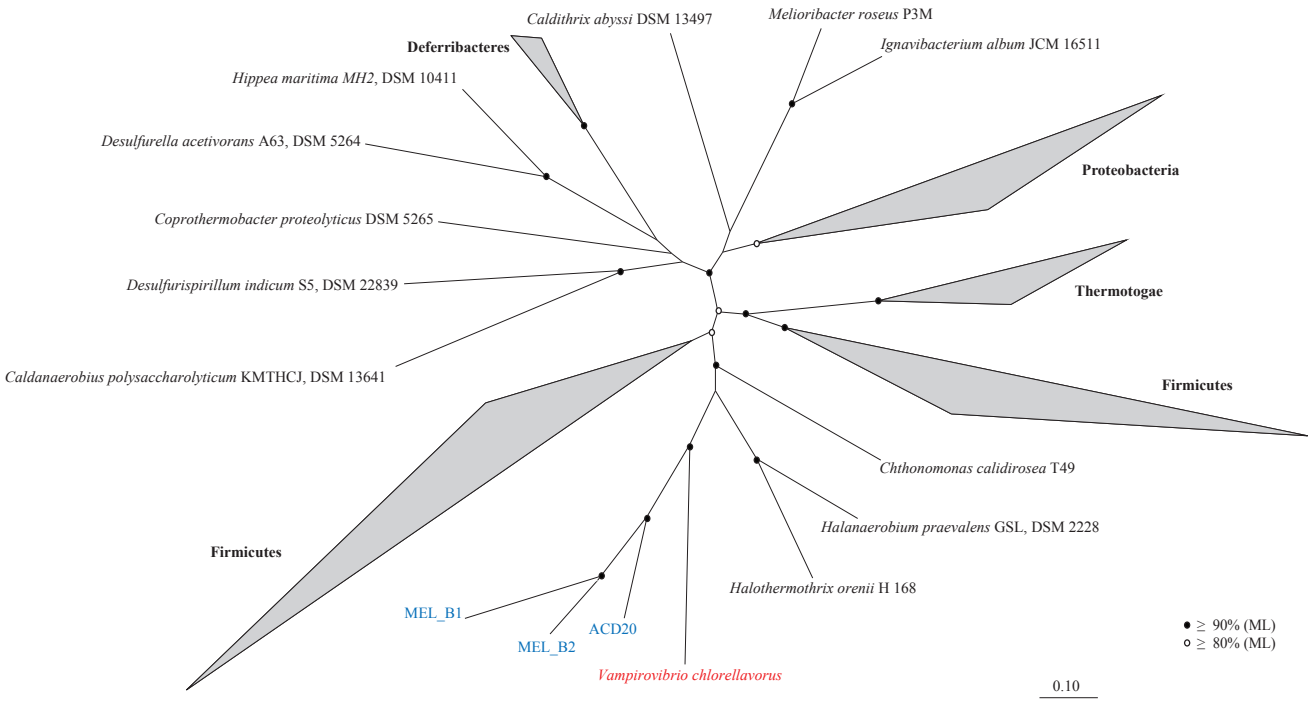


Figure S3.6. A maximum likelihood phylogenetic tree of *fliI* genes. The phylogenetic tree is constructed from 2,256 finished genomes from the IMG database (Markowitz *et al.,* 2009). The tree is unrooted and only the Melainabacteria and its closest neighbours are shown. *V. chlorellavorus* is in red and the other three Melainabacteria representatives are in blue. Phyla are in bold. Black circles in the tree represents nodes with ≥ 90% bootstrap support and white circles represents nodes with ≥ 80% bootstrap support.

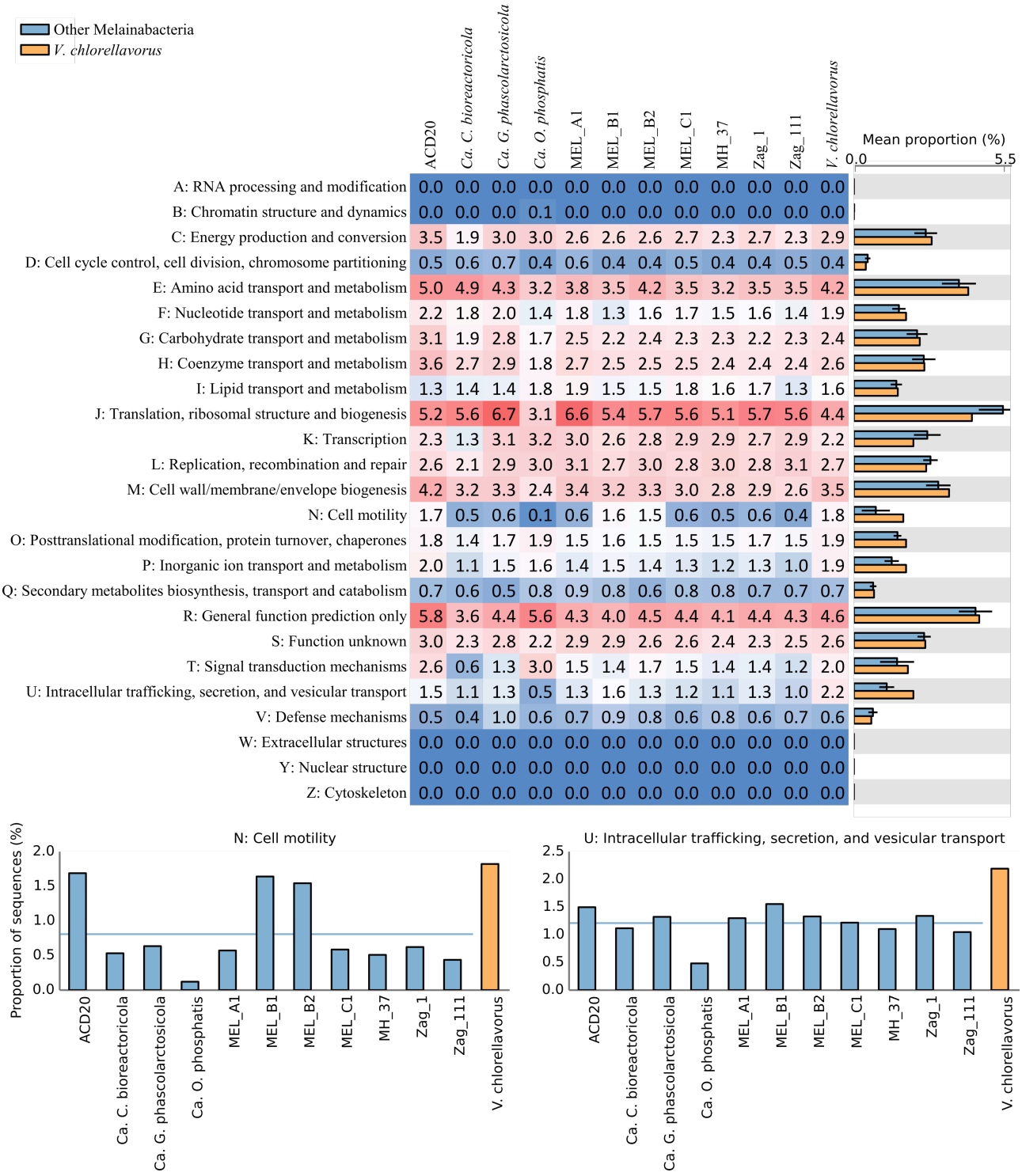


Figure S3.7. Clusters of orthologous groups (COGs) for the class Melainabacteria. There is an increase in the COGs associated with cell motility (category N) and intracellular trafficking, secretion, and vesicular transport (category U) for *V. chlorellavorus*. STAMP (Parks *et al.,* 2014) was used to explore the resulting COG profiles and create summary plots.

Supplementary Tables

Table S3.1. Marker genes used for constructing the concatenated gene tree.
A set of 178 single copy genes present exactly once in >90% of the trusted genomes (found in >90% of the genomes) from the Integrated Microbial Genomes (IMG; (Markowitz *et al.,* 2014)) database was identified. From the 178 initial genes, 69 were removed from consideration as they exhibited divergent phylogenetic histories in >1% of the trusted genomes. The remaining 109 genes were used to construct a concatenated gene tree (**Figure S3.2**).

Gene ID	Description	Length (aa)
Included markers		
TIGR03723	T6A_YgjD: tRNA threonylcarbamoyl adenosine modification protein YgjD	314
TIGR03953	rplD_bact: 50S ribosomal protein L4	188
TIGR00755	ksgA: dimethyladenosine transferase	256
TIGR00138	gidB: 16S rRNA (guanine(527)-N(7))-methyltransferase GidB	183
TIGR01953	NusA: transcription termination factor NusA	340
PF00410	Ribosomal protein S8	129
PF00380	Ribosomal protein S9/S16	121
TIGR03625	L3_bact: 50S ribosomal protein L3	202
TIGR01049	rpsJ_bact: ribosomal protein S10	99
TIGR01044	rplV_bact: ribosomal protein L22	103
TIGR00964	secE_bact: preprotein translocase, SecE subunit	57
TIGR00613	reco: DNA repair protein RecO	239
TIGR00002	S16: ribosomal protein S16	78
TIGR00001	rpmI_bact: ribosomal protein L35	63
PF13507	CobB/CobQ-like glutamine amidotransferase domain	259
TIGR03591	polynuc_phos: polyribonucleotide nucleotidyltransferase	689
TIGR03594	GTPase_EngA: ribosome-associated GTPase EngA	432
TIGR00088	trmD: tRNA (guanine(37)-N(1))-methyltransferase	233
TIGR00086	smpB: SsrA-binding protein	144
TIGR00084	ruvA: Holliday junction DNA helicase RuvA	192
TIGR00082	rbfA: ribosome-binding factor A	115
TIGR00855	L12: ribosomal protein L7/L12	125
TIGR01032	rplT_bact: ribosomal protein L20	114
TIGR00019	prfA: peptide chain release factor 1	361
TIGR00396	leuS_bact: leucine--tRNA ligase	843
TIGR00012	L29: ribosomal protein L29	56
TIGR00017	cmk: cytidylate kinase	217
TIGR00150	T6A_YjeE: tRNA threonylcarbamoyl adenosine modification protein YjeE	135
TIGR00152	TIGR00152: dephospho-CoA kinase	188
TIGR00158	L9: ribosomal protein L9	148
PF00281	Ribosomal protein L5	56
TIGR00981	rpsL_bact: ribosomal protein S12	124
TIGR00090	iojap_ybeB: iojap-like ribosome-associated protein	99

TIGR00810	secG: preprotein translocase, SecG subunit	73
TIGR00092	TIGR00092: GTP-binding protein YchF	368
TIGR00095	TIGR00095: RNA methyltransferase, RsmD family	194
TIGR00250	RNAse_H_YqgF: RNAse H domain protein, YqgF family	130
TIGR01145	ATP_synt_delta: ATP synthase F1, delta subunit	172
TIGR01029	rpsG_bact: ribosomal protein S7	154
TIGR03725	T6A_YeaZ: tRNA threonylcarbamoyl adenosine modification protein YeaZ	212
TIGR01021	rpsE_bact: ribosomal protein S5	156
TIGR01024	rplS_bact: ribosomal protein L19	114
TIGR00061	L21: ribosomal protein L21	101
TIGR00337	PyrG: CTP synthase	526
TIGR00060	L18_bact: ribosomal protein L18	114
TIGR01394	TypA_BipA: GTP-binding protein TypA/BipA	594
PF10458	Valyl tRNA synthetase tRNA binding arm	66
TIGR01393	lepA: GTP-binding protein LepA	595
TIGR00436	era: GTP-binding protein Era	270
TIGR00631	uvrb: excinuclease ABC subunit B	658
TIGR00062	L27: ribosomal protein L27	83
TIGR00634	recN: DNA repair protein RecN	563
TIGR00635	ruvB: Holliday junction DNA helicase RuvB	305
TIGR00431	TruB: tRNA pseudouridine(55) synthase	210
TIGR00472	pheT_bact: phenylalanine--tRNA ligase, beta subunit	798
TIGR00496	frr: ribosome recycling factor	176
TIGR03635	S17_bact: 30S ribosomal protein S17	72
PF01192	RNA polymerase Rpb6	57
PF02576	Uncharacterised BCR, YhbC family COG0779	141
TIGR01087	murD: UDP-N-acetylmuramoylalanine--D-glutamate ligase	441
TIGR01169	rplA_bact: ribosomal protein L1	227
TIGR00487	IF-2: translation initiation factor IF-2	587
TIGR00922	nusG: transcription termination/antitermination factor NusG	172
TIGR01164	rplP_bact: ribosomal protein L16	126
TIGR01009	rpsC_bact: ribosomal protein S3	212
TIGR00043	TIGR00043: probable rRNA maturation factor YbeY	111
PF00466	Ribosomal protein L10	100
TIGR01128	holA: DNA polymerase III, delta subunit	314
TIGR00166	S6: ribosomal protein S6	95
TIGR00416	sms: DNA repair protein RadA	454
TIGR02386	rpoC_TIGR: DNA-directed RNA polymerase, beta' subunit	1147
TIGR00344	alaS: alanine--tRNA ligase	847
TIGR00615	recR: recombination protein RecR	196
TIGR02273	16S_RimM: 16S rRNA processing protein RimM	166
TIGR00360	ComEC_N-term: ComEC/Rec2-related protein	171
TIGR03654	L6_bact: ribosomal protein L6	175
TIGR01171	rplB_bact: ribosomal protein L2	275
TIGR00959	ffh: signal recognition particle protein	428
TIGR01071	rplO_bact: ribosomal protein L15	144
TIGR00952	S15_bact: ribosomal protein S15	86

TIGR01079	rplX_bact: ribosomal protein L24	104
TIGR00116	tsf: translation elongation factor Ts	293
TIGR00059	L17: ribosomal protein L17	112
PF00673	ribosomal L5P family C-terminus	95
TIGR00593	pola: DNA polymerase I	890
TIGR00595	priA: primosomal protein N'	509
TIGR01632	L11_bact: ribosomal protein L11	140
TIGR02432	lysidine_TilS_N: tRNA(Ile)-lysidine synthetase	189
TIGR00460	fmt: methionyl-tRNA formyltransferase	315
TIGR00468	pheS: phenylalanine--tRNA ligase, alpha subunit	324
TIGR01066	rplM_bact: ribosomal protein L13	141
TIGR01067	rplN_bact: ribosomal protein L14	122
PF00276	Ribosomal protein L23	92
TIGR00422	valS: valine--tRNA ligase	863
TIGR00194	uvrC: excinuclease ABC subunit C	574
TIGR02729	Obg_CgtA: Obg family GTPase CgtA	329
TIGR00020	prfB: peptide chain release factor 2	365
PF13742	OB-fold nucleic acid binding domain	99
TIGR03632	bact_S11: 30S ribosomal protein S11	117
TIGR00029	S20: ribosomal protein S20	87
TIGR03263	guanyl_kin: guanylate kinase	180
TIGR01510	coaD_prev_kdtB: pantetheine-phosphate adenylyltransferase	155
TIGR02013	rpoB: DNA-directed RNA polymerase, beta subunit	1238
TIGR03631	bact_S13: 30S ribosomal protein S13	113
TIGR00091	TIGR00091: tRNA (guanine-N(7)-)-methyltransferase	194
TIGR01050	rpsS_bact: ribosomal protein S19	92
TIGR00188	rnpA: ribonuclease P protein component	111
TIGR00877	purD: phosphoribosylamine--glycine ligase	425
TIGR01011	rpsB_bact: ribosomal protein S2	225

Removed markers

TIGR01951	nusB: transcription antitermination factor NusB	131
TIGR00033	aroC: chorismate synthase	351
TIGR00234	tyrS: tyrosine--tRNA ligase	406
PF01416	tRNA pseudouridine synthase	105
TIGR00447	pth: peptidyl-tRNA hydrolase	188
TIGR00445	mraY: phospho-N-acetylmuramoyl-pentapeptide-transferase	321
TIGR00539	hemN_rel: putative oxygen-independent coproporphyrinogen III oxidase	361
TIGR00963	secA: preprotein translocase, SecA subunit	787
TIGR00009	L28: ribosomal protein L28	58
TIGR00459	aspS_bact: aspartate--tRNA ligase	586
TIGR00456	argS: arginine--tRNA ligase	569
TIGR00008	infA: translation initiation factor IF-1	69
TIGR00083	ribF: riboflavin biosynthesis protein RibF	290
TIGR00329	gcp_kae1: metallohydrolase, glycoprotease/Kae1 family	305
TIGR01031	rpmF_bact: ribosomal protein L32	56
TIGR01034	metK: methionine adenosyltransferase	377

TIGR00398	metG: methionine--tRNA ligase	530
TIGR00420	trmU: tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	351
TIGR00179	murB: UDP-N-acetylenolpyruvoylglucosamine reductase	290
PF04127	DNA / pantothenate metabolism flavoprotein	185
TIGR02027	rpoA: DNA-directed RNA polymerase, alpha subunit	298
TIGR00382	clpX: ATP-dependent Clp protease, ATP-binding subunit ClpX	414
TIGR00065	ftsZ: cell division protein FtsZ	353
TIGR00064	ftsY: signal recognition particle-docking protein FtsY	279
TIGR01391	dnaG: DNA primase	414
TIGR00544	lgt: prolipoprotein diacylglyceryl transferase	280
TIGR00739	yajC: preprotein translocase, YajC subunit	84
TIGR01017	rpsD_bact: ribosomal protein S4	200
TIGR02191	RNaseIII: ribonuclease III	219
PF07479	NAD-dependent glycerol-3-phosphate dehydrogenase C-terminus	149
TIGR00174	miaA: tRNA dimethylallyltransferase	288
TIGR02397	dnaX_nterm: DNA polymerase III, subunit gamma and tau	355
TIGR01082	murC: UDP-N-acetylmuramate--alanine ligase	449
TIGR01083	nth: endonuclease III	192
TIGR01085	murE: UDP-N-acetylmuramyl-tripeptide synthetase	472
TIGR00575	dnlj: DNA ligase, NAD-dependent	652
TIGR00884	guaA_Cterm: GMP synthase (glutamine-hydrolyzing), C-terminal domain	310
TIGR01162	purE: phosphoribosylaminoimidazole carboxylase, catalytic subunit	156
PF01025	GrpE	166
TIGR00233	trpS: tryptophan--tRNA ligase	328
TIGR00186	rRNA_methyl_3: RNA methyltransferase, TrmH family, group 3	240
TIGR00165	S18: ribosomal protein S18	70
TIGR00414	serS: serine--tRNA ligase	418
TIGR02075	pyrH_bact: UMP kinase	233
TIGR00418	thrS: threonine--tRNA ligase	565
TIGR00419	tim: triose-phosphate isomerase	228
TIGR00362	DnaA: chromosomal replication initiator protein DnaA	437
TIGR00888	guaA_Nterm: GMP synthase (glutamine-hydrolyzing), N-terminal domain	188
TIGR00054	TIGR00054: RIP metalloprotease RseP	421
TIGR00168	infC: translation initiation factor IF-3	165
TIGR00042	TIGR00042: non-canonical purine NTP pyrophosphatase, RdgB/HAM1 family	184
TIGR00115	tig: trigger factor	410
PF02601	Exonuclease VII, large subunit	319
TIGR00663	dnan: DNA polymerase III, beta subunit	367
PF00162	Phosphoglycerate kinase	384
TIGR01063	gyrA: DNA gyrase, A subunit	800
TIGR01060	eno: phosphopyruvate hydratase	425

TIGR02727	MTHFS_bact: 5-formyltetrahydrofolate cyclo-ligase	182
TIGR02350	prok_dnaK: chaperone protein DnaK	596
TIGR00580	mfd: transcription-repair coupling factor	923
TIGR03534	RF_mod_PrmC: protein-(glutamine-N5) methyltransferase, release factor-specific	253
TIGR02012	tigrfam_recA: protein RecA	321
TIGR00006	TIGR00006: 16S rRNA (cytosine(1402)-N(4))-methyltransferase	310
TIGR00442	hisS: histidine--tRNA ligase	406
TIGR01051	topA_bact: DNA topoisomerase I	632
TIGR00967	3a0501s007: preprotein translocase, SecY subunit	414
PF00342	Phosphoglucose isomerase	486
TIGR00643	recG: ATP-dependent DNA helicase RecG	629
TIGR00392	ileS: isoleucine--tRNA ligase	861

Table S3.2. Predicted highly expressed (PHX) and alien genes. PHX and alien gene prediction was performed with PHX analysis, using ribosomal proteins, chaperones and transcriptional and translational proteins of *V. chlorellavous* as representatives of recognised highly expressed genes to identify other putatively highly expressed genes in the genome (Karlin and Mrázek, 2000).

Highly expressed and alien gene

IMG number	IMG annotation	Eg number
2600256191	Protein-export membrane protein, SecD/SecF family	1.05

Highly expressed genes

IMG number	IMG annotation	Eg number
2600254915	ATP synthase F1 subcomplex beta subunit	1.50
2600254966	Superfamily II DNA and RNA helicases	1.10
2600254993	Parvulin-like peptidyl-prolyl isomerase	1.17
2600255027	DNA-binding protein, YbaB/EbfC family	1.20
2600255064	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta subunit	1.70
2600255065	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha subunit	1.32
2600255172	Thioredoxin-like proteins and domains	1.10
2600255235	5'-deoxy-5'-methylthioadenosine phosphorylase	1.45
2600255249	Sugar transferases involved in lipopolysaccharide synthesis	1.05
2600255283	P22 coat protein - gene protein 5	1.12
2600255291	Vacuolar-type H(+)-translocating pyrophosphatase	1.69
2600255297	Aspartyl-tRNA synthetase, bacterial type	1.35
2600255302	Hypothetical protein	1.10
2600255315	S-adenosylmethionine decarboxylase	1.28
2600255327	Predicted P-loop-containing kinase	1.01
2600255352	Chaperone protein DnaK	1.62
2600255416	NADH dehydrogenase subunit A (EC 1.6.5.3)	1.11
2600255450	Malate dehydrogenase (NAD) (EC 1.1.1.37)	1.13
2600255470	Membrane protein insertase, YidC/Oxa1 family, C-terminal domain	1.18
2600255524	YtxH-like protein	1.02
2600255533	Signal peptide peptidase SppA, 36K type	1.05
2600255566	Outer membrane protein	1.16
2600255576	Glyceraldehyde-3-phosphate dehydrogenase (NAD+) (EC 1.2.1.12)	1.32
2600255586	RNA polymerase sigma factor, sigma-70 family	1.41
2600255592	Chaperonin GroL	1.86
2600255626	Ribosomal protein L7/L12	1.43
2600255627	DNA-directed RNA polymerase subunit beta (EC 2.7.7.6)	1.41
2600255628	DNA-directed RNA polymerase gamma chain (EC 2.7.7.6)	1.49
2600255637	Two component transcriptional regulator, LuxR family	1.10
2600255664	Ribonucleotide reductase, beta subunit	1.14
2600255681	ATPases with chaperone activity, ATP-binding subunit	1.21
2600255723	Cbb3-type cytochrome oxidase, subunit 1	1.07
2600255734	Hypothetical protein	1.08
2600255835	Flagellar basal body L-ring protein	1.14

2600255847	Superoxide dismutase	1.30
2600255852	Ribosomal protein L9	1.41
2600255862	Protein of unknown function (DUF561)	1.38
2600255880	Septum site-determining protein MinD	1.07
2600255892	Ribosomal protein S3, bacterial type	1.52
2600255895	LSU ribosomal protein L2P	1.40
2600255900	Translation elongation factor 1A (EF-1A/EF-Tu)	1.41
2600255911	Translation elongation factor EF-G	2.00
2600255941	Rare lipoprotein A	1.24
2600255966	Ribosomal protein L17	1.31
2600255967	DNA-directed RNA polymerase, alpha subunit, bacterial and chloroplast-type	1.23
2600255975	LSU ribosomal protein L15P	1.34
2600255977	Ribosomal protein S5, bacterial/organelle type	1.59
2600256085	Polyribonucleotide nucleotidyltransferase	1.71
2600256086	SSU ribosomal protein S15P	1.07
2600256099	Translation elongation factor P	1.21
2600256132	Phosphopentomutase	1.21
2600256134	Uncharacterized protein conserved in bacteria	1.04
2600256170	Uncharacterized conserved protein	1.05
2600256171	6-pyruvoyl-tetrahydropterin synthase	1.11
2600256173	Bacterial regulatory proteins, gntR family	1.23
2600256207	Enoyl-[acyl-carrier-protein] reductase [NADH] (EC 1.3.1.9)	1.30
2600256221	Hypothetical protein	1.16
2600256238	Ribose-phosphate pyrophosphokinase	1.15
2600256248	Hypothetical protein	1.28
2600256265	Aconitase (EC 4.2.1.3)	1.11
2600256266	Isocitrate dehydrogenase (NADP) (EC 1.1.1.42)	1.34
2600256282	Peroxiredoxin	1.53
2600256283	Ribosomal protein L1, bacterial/chloroplast	1.47
2600256284	LSU ribosomal protein L11P	1.19
2600256314	S1 RNA binding domain	1.67
2600256336	S1 RNA binding domain	1.38
2600256348	Adenosylhomocysteinase (EC 3.3.1.1)	1.38
2600256354	Carbon storage regulator, CsrA	1.18
2600256361	NusA antitermination factor	1.15
2600256362	Translation initiation factor IF-2	1.24
2600256385	Bacterial SH3 domain	1.28
2600256386	Hypothetical protein	1.84
2600256388	Thioredoxin	1.36
2600256401	Hypothetical protein	1.08
2600256411	FKBP-type peptidyl-prolyl cis-trans isomerases 2	1.14
2600256417	Inosine-5'-monophosphate dehydrogenase (EC 1.1.1.205)	1.06
2600256436	Glycosyl hydrolases family 8	1.10
2600256437	Threonyl-tRNA synthetase (EC 6.1.1.3)	1.02
2600256466	6-phosphofructokinase	1.10
2600256529	LL-diaminopimelate aminotransferase apoenzyme (EC 2.6.1.83)	1.27
2600256549	Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit C (EC 6.3.5.-)	1.12

2600256550	CTP synthase (EC 6.3.4.2)	1.10
2600256560	Fe2+/Zn2+ uptake regulation proteins	1.11
2600256561	Rubrerythrin	1.09
2600256583	Two component transcriptional regulator, LuxR family	1.12
2600256588	Nucleoside diphosphate kinase (EC 2.7.4.6)	1.41
2600256597	Cyanobacterial porin (TC 1.B.23)	1.67
2600256623	2-oxoacid:acceptor oxidoreductase, alpha subunit	1.79
2600256624	2-oxoacid:acceptor oxidoreductase, beta subunit, pyruvate/2-ketoisovalerate family	1.41
2600256651	Flagellar hook-basal body protein	1.17
2600256660	Biopolymer transport proteins	1.08
2600256663	Hypothetical protein	1.09
2600256675	3-deoxy-D-arabinoheptulosonate-7-phosphate synthase (EC 2.5.1.54)	1.03
2600256710	GTP-binding protein TypA/BipA	1.46
2600256731	Hypothetical protein	1.13
2600256733	Adenine phosphoribosyltransferase (EC 2.4.2.7)	1.18
2600256757	ATP synthase F1 subcomplex alpha subunit	1.31
2600256779	Hemolysin activation/secretion protein	1.23
2600256821	Phosphate transport regulator	1.05
2600256833	Hypothetical protein	1.06
2600256897	YlqD protein	1.18
2600256918	Exodeoxyribonuclease VII small subunit (EC 3.1.11.6)	1.06
2600256944	S-layer homology domain	1.07
2600256966	RNA polymerase, sigma 28 subunit, SigD/FliA/WhiG	1.38
2600256998	Glutaredoxin-related protein	1.14
2600257077	Cell division protein FtsZ	1.28
2600257160	Alpha-glucan phosphorylases	1.38
2600257195	Peroxiredoxin, OsmC subfamily	1.43
2600257223	Type III secretion system ATPase, FliI/YscN	1.08
2600257249	Hypothetical protein	1.13
2600257292	Protein of unknown function (DUF1292)	1.33
2600257293	Succinyl-CoA synthetase (ADP-forming) beta subun (EC 6.2.1.5)	1.28
2600257294	Succinyl-CoA synthetase, alpha subunit	1.27
2600257316	Ribosomal protein S2, bacterial type	1.79
2600257317	Translation elongation factor Ts (EF-Ts)	1.15
2600257354	Hypothetical protein	1.12
2600257355	Hypothetical protein	1.10
2600257409	ATP-dependent Clp protease ATP-binding subunit ClpX (EC 3.4.21.92)	1.00
2600257410	ATP-dependent Clp protease proteolytic subunit ClpP (EC 3.4.21.92)	1.10
2600257411	Trigger factor	1.48
2600257427	Succinate dehydrogenase/fumarate reductase, flavoprotein subunit	1.30
2600257445	Flagellar basal body rod protein	1.07
2600257449	Flagellar basal-body rod protein FlgC	1.06

Alien genes

IMG number	IMG annotation	Eg number
2600256772	Hypothetical protein	0.99
2600255126	Hypothetical protein	0.98
2600256735	Hypothetical protein	0.92
2600255443	Hypothetical protein	0.88
2600257671	Periplasmic protein involved in polysaccharide export	0.88
2600257031	Hypothetical protein	0.86
2600257567	Imidazoleglycerol-phosphate synthase	0.86
2600256774	Beta-lactamase class C and other penicillin binding proteins	0.85
2600257020	Hypothetical protein	0.85
2600257499	Hypothetical protein	0.85
2600257685	Hypothetical protein	0.85
2600255194	Hypothetical protein	0.84
2600255922	FOG: CheY-like receiver	0.84
2600256485	prepilin-type N-terminal cleavage/methylation domain	0.84
2600257568	Imidazole glycerol phosphate synthase, glutamine amidotransferase subunit	0.84
2600256419	Hypothetical protein	0.83
2600256973	Hypothetical protein	0.82
2600257103	Outer membrane protein	0.82
2600257113	Hypothetical protein	0.82
2600257585	Transposase and inactivated derivatives	0.82
2600257699	Conjugal transfer protein TrbH	0.82
2600257732	AAA ATPase domain/AAA domain	0.82
2600255868	Hypothetical protein	0.81
2600257187	Twin arginine targeting (Tat) protein translocase TatC	0.81
2600257497	Diadenosine tetraphosphate (Ap4A) hydrolase and other HIT family hydrolases	0.81
2600257527	Nucleotidyl transferase of unknown function (DUF1814)	0.81
2600257529	Nucleotidyl transferase of unknown function (DUF1814)	0.81
2600257563	Methyltransferase domain	0.81
2600255144	Hypothetical protein	0.80
2600255224	Hypothetical protein	0.80
2600255777	TIR domain	0.80
2600257632	Uncharacterized protein conserved in bacteria	0.80
2600257676	Glycosyl transferase family 2	0.80
2600254951	Hypothetical protein	0.79
2600255590	Hypothetical protein	0.79
2600255756	Hypothetical protein	0.79
2600255862	Protein of unknown function (DUF561)	0.79
2600256409	Domain of unknown function (DUF4145)	0.79
2600256512	Alginate lyase	0.79
2600257109	RHS repeat-associated core domain	0.79
2600257362	Hypothetical protein	0.79
2600257480	Plasmid encoded RepA protein	0.79
2600257489	TaqI-like C-terminal specificity domain/ Methyltransferase domain	0.79
2600257491	Hypothetical protein	0.79
2600257570	Hypothetical protein	0.79
2600257594	Plasmid encoded RepA protein	0.79

2600257630	Predicted ATP-dependent endonuclease of the OLD family	0.79
2600257665	Hypothetical protein	0.79
2600257691	Trypsin-like peptidase domain/PDZ domain	0.79
2600255719	Hypothetical protein	0.78
2600257105	RND family efflux transporter, MFP subunit	0.78
2600257230	Dehydrogenases (flavoproteins)	0.78
2600257492	Predicted ATPase (AAA+ superfamily)	0.78
2600257669	Teichoic acid biosynthesis proteins	0.78
2600255978	Predicted phosphohydrolases	0.77
2600256525	Hypothetical protein	0.77
2600256642	Hypothetical protein	0.77
2600257225	Protein of unknown function (DUF2971)	0.77
2600257562	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis	0.77
2600257599	Domain of unknown function (DUF389)	0.77
2600257674	Hypothetical protein	0.77
2600257711	DnaA N-terminal domain	0.77
2600256983	Peptidase M15	0.76
2600257091	Hypothetical protein	0.76
2600257561	Nucleoside-diphosphate-sugar epimerases	0.76
2600257610	Uncharacterized conserved protein (COG2071)	0.76
2600257667	Nucleoside-diphosphate-sugar pyrophosphorylase involved in lipopolysaccharide biosynthesis/translation initiation factor 2B	0.76
2600257680	Endoglucanase	0.76
2600255374	Hypothetical protein	0.75
2600255506	Predicted phosphohydrolases	0.75
2600257104	The (Largely Gram-negative Bacterial) Hydrophobe/Amphiphile Efflux-1 (HAE1) Family	0.75
2600257228	Predicted naringenin-chalcone synthase	0.75
2600257508	Hypothetical protein	0.75
2600257517	Hypothetical protein	0.75
2600257663	Hypothetical protein	0.75
2600257672	Uncharacterized protein involved in exopolysaccharide biosynthesis	0.75
2600257678	Glycosyltransferase	0.75
2600255258	Dihydrofolate reductase	0.74
2600255507	Type II secretion system (T2SS), protein F	0.74
2600256033	Sugar transferases involved in lipopolysaccharide synthesis	0.74
2600256720	Hypothetical protein	0.74
2600257493	Adenine-specific DNA methylase containing a Zn-ribbon	0.74
2600257537	RND family efflux transporter, MFP subunit	0.74
2600257646	P-type conjugative transfer protein TrbL	0.74
2600255269	Hypothetical protein	0.73
2600257541	HipA-like C-terminal domain/HipA N-terminal domain/ HipA-like N-terminal domain	0.73
2600257572	ABC-type multidrug transport system, ATPase and permease components	0.73
2600257666	UDP-galactose 4-epimerase (EC 5.1.3.2)	0.73
2600257677	Glycosyltransferase	0.73
2600254929	Ankyrin repeats (3 copies)	0.72
2600255090	Restriction endonuclease S subunits	0.72

2600255508	Flp pilus assembly protein TadB	0.72
2600256159	Glycosyltransferase	0.72
2600257232	Ankyrin repeats (3 copies)/Ankyrin repeats	0.72
2600255059	Hypothetical protein	0.71
2600256873	Hypothetical protein	0.71
2600257088	UvrD-like helicase C-terminal domain/UvrD/REP helicase N-terminal domain	0.71
2600257494	Hypothetical protein	0.71
2600257607	DoxX-like family	0.71
2600257673	Lipid A core - O-antigen ligase and related enzymes	0.71
2600257675	Glycosyltransferases involved in cell wall biogenesis	0.71
2600257686	Predicted ATPase (AAA+ superfamily)	0.71
2600255060	Hypothetical protein	0.70
2600255674	Translation factor SUA5	0.70
2600257601	Trehalose-6-phosphate synthase	0.70
2600255536	Hypothetical protein	0.68
2600256155	Hypothetical protein	0.67
2600256156	Glycosyltransferase	0.69
2600257209	Ammonium transporter (TC 1.A.11)	0.69
2600257275	Hypothetical protein	0.69
2600257516	Type IV secretory pathway, VirB10 components	0.68
2600256157	Coenzyme F390 synthetase	0.68
2600256158	Coenzyme F390 synthetase	0.68
2600256160	Membrane protein involved in the export of O-antigen and teichoic acid	0.68
2600256524	Hypothetical protein	0.68
2600257575	Predicted dehydrogenase	0.68
2600257231	Predicted membrane protein	0.67
2600257569	Hypothetical protein	0.67
2600257369	Hypothetical protein	0.66
2600257564	Asparagine synthase	0.66
2600257017	Dolichyl-phosphate-mannose-protein mannosyltransferase	0.64

Table S3.3. Flagella and type IV pili genes encoded on the *V. chlorellavorus* genome. Putative gene numbers are assigned using IMG/ER (Markowitz *et al.,* 2009).

ORF number	Gene	Function
Flagella		
2600255340	*fliH*	Flagellar assembly protein
2600255341	*flhA*	Flagellar biosynthesis pathway
2600255378	*flgC*	Flagella basal body rod protein
2600255418	*motB*	Flagellar motor protein
2600255419	*motA*	Flagellar motor component
2600255528	*flgE*	flagellar hook-basal body protein
2600255530	*fliJ*	flagellar export protein
2600255622	*fliS*	flagellar biosynthetic protein
2600255623	*fliT*	Flagellar protein
2600255743	*fliP*	flagellar biosynthetic protein
2600255744	*fliQ*	flagellar biosynthetic protein
2600255745	*fliR*	flagellar biosynthetic protein
2600255746	*flhB*	flagellar biosynthetic protein
2600255747	*flhA*	flagellar biosynthesis protein
2600255829	*flgL*	flagellar hook-associated protein 3
2600255830	*flgK*	flagellar hook-associated protein
2600255831	*flgN*	protein
2600255833	*flgJ*	Rod binding protein
2600255834	*flgI*	Flagellar basal-body P-ring protein
2600255835	*flgH*	Flagellar basal body L-ring protein
2600255836	*flgA*	flagella basal body P-ring formation protein
2600255837	*flgG*	flagellar basal-body rod protein
2600255838	*flgG*	flagellar hook-basal body protein
2600256111	*flgB*	flagellar basal-body rod protein
2600256112	*flgC*	flagellar basal-body rod protein
2600256113	*fliE*	flagellar hook-basal body complex protein
2600256114	*fliF*	flagellar basal-body M-ring protein/flagellar hook-basal body protein
2600256115	*fliG*	flagellar motor switch protein
2600256116	*fliH*	Flagellar biosynthesis/type III secretory pathway protein
2600256117	*fliI*	type III secretion system ATPase (EC 3.6.3.15)
2600256210	*fliN*	flagellar motor switch protein
2600256211	*fliM*	flagellar motor switch protein
2600256212	*fliL*	Flagellar basal body-associated protein
2600256213	*motB*	Flagellar motor protein
2600256214	*motA*	Flagellar motor component
2600256257	*fliD*	Flagellar capping protein
2600256427	*flhB*	Flagellar biosynthesis pathway
2600256627	*fliR*	Flagellar biosynthesis pathway
2600256628	*fliQ*	Flagellar biosynthesis pathway
2600256629	*fliP*	flagellar biosynthetic protein
2600256630	*fliO*	Flagellar biosynthesis protein
2600256631	*spoA*	Flagellar motor switch/type III secretory pathway protein
2600256648	*fliJ*	Flagellar protein
2600256649	*fliK*	Flagellar hook-length control protein
2600256650	*flgD*	Flagellar hook capping protein - N-terminal region
2600256651	*flgE*	flagellar hook-basal body protein

2600257074 *fliO* Flagellar biosynthesis protein
2600257445 *flgF* Flagellar basal body rod protein
2600257446 *flgG* flagellar hook-basal body protein
2600257447 *flgB* Flagellar basal body protein
2600257449 *flgC* flagellar basal-body rod protein
2600257451 *fliE* flagellar hook-basal body complex protein
2600257452 *fliF* Flagellar biosynthesis/type III secretory pathway lipoprotein

---

Pili

2600255034 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255035 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255080 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255146 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255147 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255148 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255219 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255220 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255349 *pulF* Type II secretory pathway, component PulF
2600255350 *pulE/pilB* Type II secretory pathway, ATPase PulE/Tfp pilus assembly pathway, ATPase PilB
2600255395 *tadD* Flp pilus assembly protein TadD, contains TPR repeats
2600255417 *tadD* Flp pilus assembly protein TadD, contains TPR repeats
2600255465 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255474 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255501 *tadD* Flp pilus assembly protein TadD, contains TPR repeats
2600255508 *tadB* Flp pilus assembly protein TadB
2600255509 *cpaF* Flp pilus assembly protein, ATPase CpaF
2600255510 *cpaE* Flp pilus assembly protein, ATPase CpaE
2600255511 *cpaC* Flp pilus assembly protein, secretin CpaC
2600255512 *cpaB* Flp pilus assembly protein CpaB
2600255572 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255587 *pulF/pilC* Type II secretory pathway, component PulF
2600255588 *pulE/pilB* Type II secretory pathway, ATPase PulE/Tfp pilus assembly pathway, ATPase PilB
2600255589 *pulD* Type II secretory pathway, component PulD
2600255642 *pilT* pilus retraction protein PilT
2600255666 - Flp pilus assembly protein, pilin Flp
2600255863 *pilF* Tfp pilus assembly protein PilF
2600255931 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255932 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255933 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600255934 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256022 *pilF* Tfp pilus assembly protein PilF
2600256198 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256230 *pilV* Tfp pilus assembly protein PilV
2600256231 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256293 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256298 *tadE* TadE-like protein
2600256299 *cpaB* Flp pilus assembly protein CpaB
2600256300 - BON domain/Bacterial type II and III secretion system protein/Pilus formation protein N terminal region
2600256301 *cpaE* Flp pilus assembly protein, ATPase CpaE

2600256302 *cpaF* Flp pilus assembly protein, ATPase CpaF
2600256303 *tadB* Flp pilus assembly protein TadB
2600256304 *tadC* Flp pilus assembly protein TadC
2600256317 *pulO* Type II secretory pathway, prepilin signal peptidase PulO and related peptidases
2600256344 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256395 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256424 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256425 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256484 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256485 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256491 *pilT* pilus retraction protein PilT
2600256635 *tadD* Flp pilus assembly protein TadD, contains TPR repeats
2600256665 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256723 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256775 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256776 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600256849 *pilF* Tfp pilus assembly protein PilF
2600256953 *pilN* Fimbrial assembly protein (PilN)
2600256956 *pilF* Tfp pilus assembly protein PilF
2600257134 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600257156 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600257173 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600257271 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600257291 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600257309 *tadE/G* Putative Flp pilus-assembly TadE/G-like
2600257323 - Flp pilus assembly protein, pilin Flp
2600257324 - Flp pilus assembly protein, pilin Flp
2600257352 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600257353 *pulG* prepilin-type N-terminal cleavage/methylation domain
2600257403 - Flp pilus assembly protein, pilin Flp

Table S3.4. Carbohydrate-active enzymes and proteases encoded on the *V. chlorellavous* genome. Putative genes were annotated with the dbCAN web server (Yin *et al.,* 2012) to identify glycoside hydrolases and checked against the IMG annotations and BLAST results.

	Known activity	Pfam families and domains	Number
Glycoside Hydrolases			
Cellulases			
GH5	Endoglucanase	PF00150	1
GH6	Endoglucanase	PF01341	1
GH8	Endoglucanase	PF01270	2
GH9	Endoglucanase	PF02927/PF00759	1
Murein degradation (T4SS)			
GH23	Transglycosylase	PF01464	7
Oligosaccharide-degrading enzymes			
GH3	Beta-galactosidase	PF00933	1
GH35	Beta-galactosidase	PF01301	1
GH38	Alpha-mannosidase	PF01074/PF09261	1
Polysaccharide-degrading enzymes			
GH13	Mainly α-amylase	PF00128/PF02638/PF02806/PF02922/PF11941	7
GH77	4-α-glucanotransferase	PF02446	1
Other			
GH4	Glycerol-3-phosphate dehydrogenase	PF01210/PF07479	1
GH15	Glucoamylase?	PF00723	1
GH19	Chitinase	PF01471	1
GH57	α-amylase, 4-α-glucanotransferase	PF03065/PF09094	2
GH109	Oxidoreductase, 3-hydroxyisobutyrate dehydrogenase, saccharopine dehydrogenase	PF01408/PF02894/PF03435/PF03446/PF14833	6
Glycosyltransferases			
GT1	-	PF13528	1
GT2	-	PF00534/PF00535/PF07238/PF13579/PF13641	18
GT4	Kdotransferase	PF04413/PF00534/PF00535/PF13439/PF13579/ PF13477	17
GT5	-	PF08323/PF00534	3

GT9	Heptosyltransferase	PF01075	5
GT19	Lipid-A-disaccharide synthetase/ UDP-N-acetylglucosamine 2-epimerase	PF02684/PF02350	2
GT20	-	PF00982	1
GT26	Glycosyl transferase WecB/TagA/CpsF	PF02706/PF03808	5
GT27	-	PF13641	1
GT28	Monogalactosyldiacylglycerol synthase	PF03033/PF06925/PF13528	3
GT30	-	PF00534/PF04413	1
GT35	Phosphorylase	PF00343	2
GT39	Dolichyl-phosphate-mannose-protein Mannosyltransferase	PF13231	1
GT51	Transglycosylase/Transpeptidase	PF00905/PF00912	1
GT83	Dolichyl-phosphate-mannose-protein Mannosyltransferase	PF13231	6
Auxillary Activities			
AA2	Catalase/peroxidase	PF00141	3
AA6	Flavoprotein	PF03358	1
AA7	FAD/FMN-containing dehydrogenase	PF01565/PF02913	2
Carbohydrate-Binding Modules			
CBM48	1,4-alpha-glucan branching enzyme	PF00128/PF02922/PF11941	2
CBM50	LysM domain	PF01476	1
CBM57	Melectin	PF11721	2
Carbohydrate Esterases			
CE4	Polysaccharide deacetylase	PF01522	1
CE7	Esterase/lipase/hydrolase	PF12695	3
CE9	N-acetylglucosamine-6-phosphate deacetylase/cytosine deaminase	PF01979	2
CE10	Esterase/lipase	PF12697/PF12695	2
CE11	UDP-3-O-acyl-N-acetylglucosamine deacetylase	PF03331	1

Polysaccharide Lyases			
PL7	Alginate lyase	PF08787	1

Table S3.5. Peptidases encoded on the *V. chlorellavorus* genome. The MEROPS server (Rawlings *et al.,* 2014) was used to identify putative peptidases in *V. chlorellavorus* using batch BLAST.

Aspartic	
Cytoplasmic	2
Cytoplasmic membrane	1
Unknown	1
Cysteine	
Cytoplasmic	9
Unknown	5
Serine	
Cytoplasmic	11
Cytoplasmic membrane	5
Periplasmic	5
Extracellular	1
Unknown	15
Metallo	
Cytoplasmic	21
Cytoplasmic membrane	7
Periplasmic	3
Outer membrane	2
Unknown	12
Inhibitor	
Cytoplasmic	1
Outer membrane	1
NB	
Cytoplasmic	2
Unknown	
Cytoplasmic	1

Appendix C: Supplementary figures and tables for Chapter 4

Supplementary Figure

Figure S4.1. Sample P3D
Collected from the active layer palsa (30-33cm below the surface) in Stordalen Mire, northern Sweden in July 2012.

Supplementary Tables

Table S4.1. Genes belong to pathways from the *Obscuribacterales* representatives from this study

Numbers in the far right corner correspond to the numbers in **Figures 4.2** and **4.3**. Colours refer to expression levels based on metatranscriptomics data. Orange (FPKG <0-20), yellow (FPKG 21-50), green (FPKG 51-100), blue (FPKG 101-500), purple (501-1000), red (FPKG >1000).

	EC #	Gene	P3DObs1 IMG accession number	P3DObs2 IMG accession number	
Cell wall and shape					
Outer membrane					
Nucleoside-diphosphate-sugar pyrophosphorylase	2.7.7.13	-	2606134928	2606140980	
			2606137375	2606143460	
Sugar transferases involved in lipopolysaccharide synthesis	-	*wbqP*	2606136182	2606140635	
			2606137759	2606143850	
Peptidoglycan/LPS O-acetylase OafA/YrhL, contains acyltransferase and SGNH-hydrolase domains	-	-	2606131800	2606139417	
			2606134965	2606140417	
			2606135228	2606141600	
Muramoyltetrapeptide carboxypeptidase LdcA (peptidoglycan recycling)			2606136763	2606139344	
			2606136764	2606139345	
ADP-heptose:LPS heptosyltransferase	-	-	2606132132	-	
			2606133331	-	
			2606133399	-	
UDP-N-acetylglucosamine:LPS N-acetylglucosamine transferase	2.4.1.157	*ugtP*	2606134349	2606143975	
			2606136249	-	
O-antigen ligase	-	-	2606132673	2606140330	
			2606133537	2606140558	
O-antigen ligase like membrane protein	-	-	2606137741		
Membrane protein involved in the export of O-antigen and teichoic acid	-	-	2606133358	2606139121	
			2606135200	2606142531	
			2606137780		
Lipopolysaccharide biosynthesis protein, LPS:glycosyltransferase	-	-	-	2606140171	
Lipid-A-disaccharide kinase	2.4.1.182	*ipxK*	2606133332	2606142893	
			2606136064	2606143678	
Rod shape					
Rod shape-determining protein MreB	-	*mreB*	2606133294	2606142856	
Rod shape-determining protein MreC	-	*mreC*	2606133293	2606142855	
Rod shape-determining protein MreD	-	*mreD*	2606133292	2606142854	
Penicillin-binding protein 2	-	*mrdA*	2606133291	2606142853	
Rod shape-determining protein RodA	-	*rodA*	-	2606139744	
Cytoskeletal protein RodZ	-	*rodZ*	2606133272	-	
Energy metabolism					
TCA cycle					

Phosphoenolpyruvate carboxykinase (GTP)	4.1.1.32	*pckA*	2606132530	2606140390	
			2606133597	2606142622	
			2606134464	-	
			2606134991	-	
Citrate synthase	2.3.3.1	*gltA*	2606137359	2606140963	1
			2606137360	2606140965	2
			-	2606138655	3
			-	2606141706	4
Aconitase	4.2.1.3	*acnA*	2606136701	-	5
Isocitrate dehydrogenase (NADP)	1.1.1.42	*icd*	-	-	
2-oxoacid:acceptor oxidoreductase, alpha subunit	1.2.7.3	*korA*	2606133448	2606139035	6
			-	2606139390	
2-oxoacid:acceptor oxidoreductase, beta subunit, pyruvate/2-ketoisovalerate family	1.2.7.3	*korB*	2606133162	2606139389	7
			2606133447	2606139036	7
Dihydrolipoamide dehydrogenase	1.8.1.4	*pdhD*	2606132243	2606139367	8
2-oxoglutarate dehydrogenase, E1 component	1.2.4.2	*sucA*	2606137480	-	9
Succinyl-CoA synthetase, alpha subunit	6.2.1.5	*sucD*	2606131698	2606138825	10
Succinyl-CoA synthetase, beta subunit	6.2.1.5	*sucC*	2606131699	2606138826	11
Succinate dehydrogenase/fumarate reductase, flavoprotein subunit	1.3.5.1	*sdhB*	2606136695	-	12
Succinate dehydrogenase and fumarate reductase iron-sulfur protein	1.3.5.1	*sdhA*	2606136696	-	13
Fumarase, class II	4.2.1.2	*fumC*	2606131843	2606138495	14
			2606133434	2606139048	14
Malate dehydrogenase (NAD)	1.1.1.37	*mdh*	2606137837	2606141647	15
Pyruvate dehydrogenase E1 component, alpha subunit	1.2.4.1	*pdhA*	2606132246	2606139364	16
Pyruvate/2-oxoglutarate dehydrogenase complex, dehydrogenase (E1) component, eukaryotic type, beta subunit	1.2.4.1	*pdhB*	2606132245	2606139365	17
Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide acyltransferase (E2) component, and related enzymes	2.3.1.12	*pdhC*	2606132244	2606139366	18
Dihydrolipoamide dehydrogenase	1.8.1.4	*pdhD*	2606132243	2606139367	19
Glyoxylate shunt					
Isocitrate lyase	4.1.3.1	*aceA*	2606132916	2606140966	20
			2606137361	2606140984	20
Malate synthase	2.3.3.9	*aceB*	2606136724	-	21
Electron transport chain					
NADH dehydrogenase (Complex I)					
NADH:ubiquinone oxidoreductase subunit 3 (chain A)	1.6.5.3	*nuoA*	2606134554	2606142182	22
Respiratory-chain NADH dehydrogenase, 30 Kd subunit	1.6.5.3	*nuoC*	2606134552	2606142180	23
NADH:ubiquinone oxidoreductase 49 kD subunit (chain D)	1.6.5.3	*nuoD*	2606134551	2606142179	24
NADH:ubiquinone oxidoreductase subunit 1 (chain H)	1.6.5.3	*nuoH*	2606134550	2606142178	25
NADH-quinone oxidoreductase, chain I	1.6.5.3	*nuoI*	2606134549	2606142177	26
Proton-translocating NADH-quinone oxidoreductase, chain M	1.6.5.3	*nuoM*	2606133212	2606139393	27
NADH dehydrogenase subunit N	1.6.5.3	*ndhB*	2606132447	2606139598	28
NADH dehydrogenase subunit L	1.6.5.3	*ndhF*	-	2606139392	29
NADH:ubiquinone oxidoreductase subunit 6 (chain J)	1.6.5.3	*ndhG*	2606134548	2606142176	30
NADH dehydrogenase subunit B	1.6.5.3	*ndhK*	2606134553	2606142181	31
NADH dehydrogenase, FAD-containing subunit	1.6.99.3	*ndh*	-	2606139632	
Succinate dehydrogenase/fumarate reductase (Complex II)					
Succinate dehydrogenase/fumarate reductase,	1.3.5.1	*sdhA*	2606136696	-	32

flavoprotein subunit					
Succinate dehydrogenase and fumarate reductase iron-sulfur protein	1.3.5.1	*sdhB*	2606136695	-	33
Cytochrome bc1 complex (Complex III)					
Cytochrome b subunit of the bc complex	-	-	2606134688	2606140315	34
				2606140318	34
Rieske Fe-S protein			2606134689	2606140314	35
Cytochrome C oxidase, cbb3-type, subunit III/Cytochrome c			2606134690	2606140313	36
Cytochrome c oxidase (Complex IV)					
Bacterioferritin (cytochrome b1)	-	*bfr*	2606133850	2606139614	
			2606135504		
Cbb3-type cytochrome oxidase, subunit 1	1.9.3.1	*ccoN*	2606134694	2606140309	37
				2606139998	37
				2606140790	37
Cytochrome c oxidase cbb3-type subunit II	-	*ccoO*	2606134693	2606140310	38
Cytochrome c oxidase, cbb3-type, subunit III		*ccoP*	2606135529	2606138074	39
				2606140226	39
Cytochrome c-type biogenesis protein CcmH/NrfG	-	*ccmH / nrfG*	2606132744	2606140744	
		-	2606133083	-	
ABC-type transport system involved in cytochrome c biogenesis, permease component	-	*ccmB*	2606137109	-	
Cytochrome c	-		2606137281	2606140225	
			2606137873	2606140808	
			2606135065	2606140810	
			2606135349	2606140812	
Cytochrome c biogenesis protein ResB	-	*resB*	2606135519	2606138064	
Cytochrome c-type biogenesis protein CcsB	-	*ccsB*	2606135520	2606138065	
Cytochrome P450	1.14.-.-	-	2606136979	2606143604	
Oxidoreductase molybdopterin binding domain/Prokaryotic cytochrome b561	-	-	-	2606139530	
Cytochrome bd complex (Complex IV)					
Cytochrome bd-type quinol oxidase, subunit 1	1.10.3.-	*cydA*	2606137370	2606140975	40
Cytochrome bd-I ubiquinol oxidase subunit 2 apoprotein	1.10.3.-	*cydB*	2606137371	2606140976	41
F-type ATPase (Complex V)					
Proton translocating ATP synthase, F1 alpha subunit	3.6.3.14	*atpA*	2606135994	2606143259	42
ATP synthase F1 subcomplex beta subunit	3.6.3.14	*atpD*	2606133762	2606143594	43
ATP synthase F1 subcomplex gamma subunit	3.6.3.14	*atpG*	2606135995	2606143258	44
ATP synthase, F1 delta subunit	3.6.3.14	*atpH*	2606135993	2606143260	45
ATP synthase F1 subcomplex epsilon subunit	3.6.3.14	*atpC*	2606133760	2606143596	46
ATP synthase F0 subcomplex A subunit	3.6.3.14	*atpB*	2606135989	2606143264	47
FoF1-type ATP synthase, membrane subunit b or b'	-	-	2606135991	2606143261	48
	3.6.3.14	*atpF*	2606135992	2606143262	48
ATP synthase, F0 subunit c	-	*atpE*	2606135990	2606143263	49
Plasma-membrane proton-efflux P-type ATPase	3.6.3.6	-	2606132974	-	
Vacuolar-type H(+)-translocating pyrophosphatase	3.6.1.1	*hppA*	2606137861	2606142270	
Nitrate reductase					
Nitrate oxidoreductase alpha subunit	1.7.99.4	*narG*	-	2606142666	50
Respiratory nitrate reductase beta subunit	1.7.99.4	*narH*	-	2606142667	51
Respiratory nitrate reductase chaperone NarJ delta subunit	-	*narJ*	-	2606142668	52
Respiratory nitrate reductase, gamma subunit	-	*narI*	-	2606142669	53
Nitrate/nitrite transporter NarK	-	*narK*	-	2606142673	
Ethanol and ethanolamine utilisation					
Ethanol utilisation					
Acyl-coenzyme A synthetase/AMP-(fatty) acid ligase	6.2.1.1	*acs*	2606134001	2606143322	54

			2606135413	2606143919	54
Aldehyde dehydrogenase	1.2.1.3	-	2606134117	2606143598	55
Alcohol dehydrogenase, class IV	1.1.1.1	*adh*	2606135499	2606141630	56
			2606136099		56
Ethanolamine utilisation					
Ethanolamine utilization protein EutA, possible chaperonin protecting lyase from inhibition	-	*eutA*	2606134618	2606141992	57
Ethanolamine ammonia-lyase light chain	4.3.1.7	*eutC*	2606134619	2606141991	58
Ethanolamine ammonia-lyase heavy chain	4.3.1.7	*eutB*	2606134620	2606141990	59
Carboxysome shell and ethanolamine utilization microcompartment protein CcmL/EutN	-	*CcmL /EutN*	2606137163	-	
NiFe hydrogenase					
Ni,Fe-hydrogenase I small subunit	-	*hyaA, hybO*	2606135125	-	60
Ni,Fe-hydrogenase I large subunit	-	*hyaB/ hybC*	2606135124	-	61
hydrogenase maturation protease	-	*hybD/ hycI*	2606135122	-	62
Pyruvate formate lyase-activating enzyme 1			2606136894	2606141677	63
Ni,Fe-hydrogenase III small subunit	-	*HydA /HypF*	2606132957	2606139229	64
			2606134507	2606141531	64
Ni,Fe-hydrogenase III large subunit	-	*HydB*	2606132956	2606139230	65
			2606134508	2606141532	65
Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MnhD subunit	-	*hyfF*	2606132955	2606140262	66
			2606134509	2606139231	66
				2606141533	66
Hydrogenase-4 membrane subunit HyfE	-	*hyfE*	2606132954	2606139232	67
			2606134510	2606140261	67
				2606141534	67
Formate hydrogenlyase subunit 4	-	*hycD*	2606132953	2606139233	68
			2606134511	2606140260	68
				2606141535	68
Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MnhD subunit		*hycC*	2606132952	2606139234	69
			2606134512	2606140259	69
				2606141536	69
Formate dehydrogenase family accessory protein FdhD		*fdhD*	2606132065	2606139611	70
Formate dehydrogenase H	1.1.99.33		2606132066	2606139612	71
NAD(P)-dependent nickel-iron dehydrogenase diaphorase component subunit HoxE	1.6.5.3	*hoxE*	2606135076	2606140844	
NAD(P)-dependent nickel-iron dehydrogenase flavin-containing subunit	1.6.5.3	*hoxF*	2606135077	2606140845	
NAD(P)-dependent nickel-iron dehydrogenase diaphorase component subunit HoxU	1.6.5.3	*hoxU*	2606135078	2606140846	
NAD(P)-dependent nickel-iron dehydrogenase subunit HoxY	1.12.1.2	*hoxY*	2606135079	2606140847	
NAD(P)-dependent nickel-iron dehydrogenase catalytic subunit	1.12.1.2	*hoxH*	2606135080	2606140848	
Zn finger protein HypA/HybF (possibly regulating hydrogenase expression)	-	*hypA/ hybF*	2606135136	2606140897	
Hydrogenase accessory protein HypB	-	*hypB*	2606135137	2606140898	
Hydrogenase maturation protein HypC	-	*hypC*	2606135132	2606140893	
Hydrogenase maturation protein HypD	-	*hypD*	2606135134	2606140895	
Hydrogenase expression/formation protein HypE	-	*hypE*	2606135135	2606140896	
[NiFe] hydrogenase maturation protein HypF	-	*hypF*	2606135131	2606140892	

Carbohydrate degradation					
Xylose degradation					
Xylose isomerase	5.3.1.5	*xylA*	2606131857	2606138469	72
			-	2606141645	72
Xylulokinase	2.7.1.17	*xylB*	2606131859	2606138467	73
Glucose degradation					
Glucokinase	2.7.1.2	*glk*	2606131371	2606138118	74
Fructose degradation					
Fructokinase	2.7.1.4	*scrK*	2606136097	2606141632	75
			-	2606142772	75
Galactose degradation					
Galactose-1-epimerase	5.1.3.3	*galM*	2606131896	2606139954	76
Galactokinase	2.7.1.6	*galK*	2606134799	2606138011	77
UDP-glucose:alpha-D-galactose-1-phosphate uridylyltransferase	2.7.7.12	*galT*	2606134798	2606138010	78
Beta-phosphoglucomutase	5.4.2.6/ 5.4.2.2	*pgm*	2606132060	2606143019	79
			-	2606143084	79
Starch degradation					
Alpha amylase	3.2.1.1	*treS*	2606131439	2606138189	80
Glycogen/starch/alpha-glucan phosphorylases	2.4.1.1	*glgP*	2606131959	2606139891	81
			2606132980	-	81
Glycogen degradation					
Glycogen phosphorylase	2.4.1.1	*glgP*	2606131959	2606139891	82
			2606132980	-	82
4-alpha-glucanotransferase	2.4.1.25	*malQ*	2606131738	2606138379	83
			2606133181	2606139370	83
Amylo-alpha-1,6,glucosidase	3.2.1.33	-	2606135852	2606138200	84
			2606136218	2606138534	84
			2606137046	2606144067	84
Glucoamylase	3.2.1.3	-	2606134790	2606138002	85
Cellulose degradation					
Cellulase	3.2.1.4	-	2606135865	2606139848	86
Broad-specificity cellobiase	3.2.1.21	-	2606133230	2606139407	87
			-	2606143727	87
Beta-galactosidase	3.2.1.21	*bglB*	2606132618	2606141143	88
			2606136402	2606139531	88
			2606134581	-	88
			2606132438	-	88
			2606133612	-	88
Mannose degradation					
Mannose-6-phosphate isomerase, class I	-	-	-	2606139513	89
Mannose-6-phosphate isomerase, cupin superfamily	-	-	-	2606142419	89
Embden-Mayerhof-Parnas pathway					
Phosphoglucomutase	5.4.2.2	*pgm*	2606132060	2606143019	90
			2606136458	2606143084	90
Glucose-6-phosphate isomerase	5.3.1.9	*pgi*	2606131369	2606138117	91
6-phosphofructokinase	2.7.1.11	*pfkA*	2606131700	2606138827	92
			2606137597	2606142930	92
Fructose-bisphosphate aldolase	4.1.2.13	*fbaB*	2606131825	2606138508	93
			2606132529	-	93
Triosephosphate isomerase	5.3.1.1	*tpiA*	2606132976	2606141638	94
			2606137848	-	94
Glyceraldehyde-3-phosphate dehydrogenase (NAD+)	1.2.1.12	*gapA*	-	-	
Phosphoglycerate kinase	2.7.2.3	*pgk*	-	-	
Phosphoglycerate mutase	5.4.2.12	*gpml*	2606131661	2606138791	95
Phosphopyruvate hydratase/enolase	4.2.1.11	*eno*	-	2606143943	96
			-	2606143769	96

Pyruvate kinase	2.7.1.40	*pyk*	2606132287	2606143438	97
			2606135182	-	97
Pentose phosphate pathway					
Ribose-5-phosphate isomerase	5.3.1.6	*rpiB*	2606132695	2606139768	98
			-	2606140578	98
Ribulose-5-phosphate 3-epimerase	5.1.3.1	*rpE*	2606138808	2606138808	99
Transketolase	2.2.1.1	*tktA*	2606131732	2606142740	100
			2606131733	2606142741	100
			2606136133	2606144004	100
Transaldolase	2.2.1.2	*tal*	2606131369	2606138117	101
			2606133170	2606139382	101
			2606136629	2606141553	101
Ribose-phosphate pyrophosphokinase	2.7.6.1	*prsA*	2606132774	2606140777	102
			2606136427	2606142043	102
Amino acid biosynthesis					
Alanine					
Cysteine desulfurase	2.8.1.7	*sufS*	2606134958	2606143650	103
Arginine					
N-acetylglutamate synthase	2.3.1.1	*argA*	2606131338	2606138087	104
			2606133015	2606141177	104
Acetylglutamate kinase	2.7.2.8	*argB*	2606134449	2606142602	105
N-acetylglutamylphosphate reductase	1.2.1.38	*argC*	2606134448	2606142601	106
Acetylornithine aminotransferase	2.6.1.11	*argD*	2606136933	2606140457	107
Acetylornithine deacetylase	3.5.1.16	-	2606133305	2606142866	108
			2606135832	2606139806	108
			2606137118	-	108
Ornithine carbamoyltransferase	2.1.3.3	*argF*	2606131465	2606138217	109
Argininosuccinate synthase	6.3.4.5	*argG*	2606135730	2606141452	110
Argininosuccinate lyase	4.3.2.1	*argH*	2606135731	2606141453	111
Asparagine					
Asparagine synthetase [glutamine-hydrolyzing] 1	6.3.5.4	*asnB*	-	2606139676	112
Aspartyl-tRNA synthetase	6.1.1.-	*aspS*	2606136882	-	113
Glutamyl-tRNA amidotransferase	6.3.5.6	*gatA*	2606136015	-	114
Aspartate					
Aspartate transaminase	2.6.1.1	*aspC*	2606132171	2606139449	115
			2606133898	2606142008	115
			2606134127	2606142019	115
			2606134605	2606142068	115
			2606136466	2606143072	115
Cysteine					
Serine acetyltransferase	2.3.1.30	*cysE*	-	2606143589	116
Cysteine synthase	2.5.1.47	*cysK*	2606133766	2606143590	117
			2606136421	2606142050	117
Cystathionine beta-synthase	4.2.1.22	-	2606133780	2606138348	118
Cystathionine gamma-lyase	4.4.1.1	-	2606133781	2606138347	119
Glutamate					
Glutamate synthase [NADPH] small chain	1.4.1.13	*gltD*	2606134596	-	120
Glutamate synthase (ferredoxin) (EC 1.4.7.1)	1.4.7.1	*gltS*	2606131467	2606138219	121
Glutamate dehydrogenase	1.4.1.3	*gdhA*	2606133985	2606143935	122
			2606134071	2606143953	122
			2606136958	2606138568	122
Glutamine					
Glutamine synthetase	6.3.1.2	*glnA*	2606133242	2606139419	123
Glycine					
Glycine dehydrogenase	1.4.4.2	-	-	2606140307	124
Aminomethyltransferase	2.1.2.10	*gcvT*	-	2606140201	125
Dihydrolipoyl dehydrogenase	1.8.1.4	-	2606132243	2606139367	126
L-threonine aldolase	4.1.2.5	*ItaE*	2606136029	2606143233	127

Proline					
Glutamate 5-kinase	2.7.2.11	*proB*	2606132831	2606138712	128
Glutamate-5-semialdehyde dehydrogenase	1.2.1.41	*proA*	2606132830	2606138711	129
Pyrroline-5-carboxylate reductase	1.5.1.2	*proC*	2606132829	2606138710	130
Threonine					
Aspartokinase	2.7.2.4	*lysC*	2606132057	2606141659	131
			2606134720	2606143017	131
			2606136876	2606139318	131
			2606137293	2606140280	131
			2606137674	-	131
Aspartate semialdehyde dehydrogenase	1.2.1.11	-	-	2606142697	132
Homoserine dehydrogenase	1.1.1.3	-	2606132054	2606143014	133
Homoserine kinase	2.7.1.39	*thrB*	2606137078	2606143016	134
			2606131854	2606138481	134
			2606132056	-	134
Threonine synthase	4.2.3.1	*thrC*	2606132055	2606143015	135
Tryptophan					
Anthranilate synthase	4.1.3.27	*trpE*	2606137232	-	136
			2606137233	-	136
Anthranilate phophoribosyl transferase	2.4.2.18	*trpD*	2606137234	-	137
Phosphoribosylanthranilate isomerase	5.3.1.24	*trpF*	2606137236	-	138
Indole-3-glycerol phosphate synthase	4.1.1.48	*trpC*	2606137235	-	139
Tryptophan synthase	4.2.1.20	*trpA*	2606137238	-	140
		trpB	2606137237	-	140
Amino acid transporters					
ABC-type amino acid transport system, permease component			2606134576		
ABC-type amino acid transport/signal transduction system, periplasmic component/domain			2606134146	2606141471	
			2606134578	-	
Amino acid transporter			2606133600	2606139329	
			2606134988	2606139740	
			2606137652	2606140394	
			-	2606143864	
			-	2606144041	
Amino acid/peptide transporter (Peptide:H+ symporter), bacterial			2606137333	2606143702	
Amino acid/polyamine/organocation transporter, APC superfamily (TC 2.A.3)			2606133135	2606138627	
			2606134238	2606140928	
			2606134393	2606142201	
ABC-type polar amino acid transport system, ATPase component			2606134577	-	
Sulphate activation pathway					
Sulfate adenylyltransferase	2.7.7.4	*cysD*	2606136165	2606143561	141
Adenylylsulfate kinase	2.7.1.25	*cysC*	2606136164	2606140614	142
2'-phospho-adenylylsulfate reductase	1.8.4.8	*cysH*	2606137072	-	143
Sulfite reductase	1.8.1.2	*cysI*	2606137068	-	144
Amino acid degradation					
Asparagine					
L-asparaginase	3.5.1.1	-	2606133321	2606142882	145
Aspartate					
Aspartate transaminase	2.6.1.1	*aspC*	2606132171	2606139449	146
			2606133898	2606142008	146
			2606134127	2606142019	146
			2606134605	2606142068	146
			2606136466	2606143072	146
Malate dehydrogenase	1.1.1.37	*mdh*	2606137837	2606141647	147

Glutamate					
Glutamate dehydrogenase	1.4.1.3	*gdhA*	2606133985	2606143935	148
			2606134071	2606143953	148
			2606136958	2606138568	148
Aspartate transaminase	2.6.1.1	*aspC*	2606132171	2606139449	149
			2606133898	2606142008	149
			2606134127	2606142019	149
			2606134605	2606142068	149
			2606136466	2606143072	149
Aspartate ammonia-lyase	4.3.1.1	*aspA*	-	2606138892	150
Glutamine					
Glutaminase	3.5.1.2	*glsA*	-	2606142426	151
glutamate synthase (NADPH) small subunit	1.4.1.13	*gltD*	2606134596	-	152
Histidine					
Histidine ammonia-lyase	4.3.1.3	*hutH*	2606135948	2606138223	153
			2606137252	-	153
			2606131472	-	153
Urocanate hydratase	4.2.1.49	*hutU*	2606134718	2606140283	154
Imidazolone-5-propionate hydrolase	3.5.2.7	*hutI*	2606131708	2606142709	155
glutamate formiminotransferase	2.1.2.5	-	2606133986	2606143934	156
Proline					
Proline dehydrogenase	1.5.5.2	-	2606132255	2606139357	157
1-pyrroline-5-carboxylate dehydrogenase	1.2.1.88	-	2606135526	2606138071	158
Tryptophan					
Tryptophan 2,3-dioxygenase apoenzyme	1.13.11.11	-	2606132488	2606139874	159
Kynurenine formamidase	3.5.1.9	*kynB*	2606133322	2606142883	160
			2606133542	2606140326	160
Kynureninase	3.7.1.3	*kynU*	2606133657	2606143048	161
Tryptophanase	4.1.99.1	-	-	2606142707	162
Serine					
Serine deaminase	4.3.1.17	-	2606136673	-	163
Alanine					
Alanine dehydrogenase	1.4.1.1	-	-	2606138887	164
Potential adaptations to a cold climate					
Sigma factors					
RNA polymerase, sigma 70 subunit, RpoD		*rpoD*	2606132677	2606140562	
			2606132936		
			2606137619		
RNA polymerase, sigma-24 subunit, RpoE		*rpoE*	2606134775		
Chaperones and stress proteins					
ATP-dependent chaperone ClpB		*clpB*	2606131490	2606138306	
Chaperone protein DnaJ		*dnaJ*	2606135408	2606141726	
			2606137398	2606141768	
				2606143327	
Molecular chaperone DnaK (HSP70)		*dnaK*	2606135407	2606139710	
				2606139722	
				2606143328	
Molecular chaperone GrpE (heat shock protein)		*grpE*	2606135406	2606143330	
Co-chaperonin GroES (HSP10)		*GroES*		2606143456	
Chaperonin GroL		*GroL*		2606143455	
Nucleotide-binding universal stress protein, UspA family		*uspA*	2606132021	2606138377	
			2606132176	2606139334	
			2606133969	2606139695	
			2606135056	2606140183	
			2606135060	2606140189	
			2606135061	2606140223	
			2606135062	2606140227	

			2606135070	2606140232	
			2606135071	2606140799	
			2606135072	2606140805	
			2606135073	2606140806	
			2606135090	2606140823	
			2606135471	2606140827	
			2606136491	2606140828	
			2606136781	2606140832	
			2606137287	2606140833	
			2606137863	2606140836	
			2606137864	2606140839	
			2606137867	2606140840	
			2606137876	2606140841	
				2606140856	
				2606140857	
				2606141024	
				2606141946	
				2606142074	
				2606142988	
				2606143874	
Transcription and translation					
Helicase conserved C-terminal domain/DEAD/DEAH box helicase/DSHCT (NUC185) domain			2606136910	2606141568	
			2606137316	2606141694	
NusA antitermination factor			2606134737	2606140486	
NusB antitermination factor			2606133529	2606140338	
			2606133998	2606143923	
Translation initiation factor 2 (bIF-2)			2606133982	2606140488	
			2606134739	2606143938	
Translation initiation factor IF-3			2606132636	2606140526	
			2606132637	-	
Translation elongation factor TU			2606137695	2606139187	
Carbon and energy reserves					
Polyphosphate kinase 1		*ppk1*	2606136248	2606142916	
Polyphosphate kinase 2, PPK2 family		*ppk2*	2606133068	2606139751	
				2606141229	
Exopolyphosphatase		*ppx*	2606133411	2606139071	
Polyphosphate:nucleotide phosphotransferase, PPK2 family			2606132977	2606138340	
			2606133789	-	
Malto-oligosyltrehalose trehalohydrolase	3.2.1.141	*treZ*	2606131432	2606138182	
Malto-oligosyltrahalose synthase	5.4.99.15	*treY*	2606131433	2606138183	
Isoamylase	3.2.1.-	*treX*	2606131434	2606138184	
			2606131445	2606142598	
Pullulanase/glycogen debranching enzyme	3.2.1.-		2606137092	-	
Cryoprotectants					
ABC-type proline/glycine betaine transport system, ATPase component			-	2606139681	
Periplasmic glycine betaine/choline-binding (lipo)protein of an ABC-type transport system (osmoprotectant binding protein)			-	2606139682	
Choline dehydrogenase or related flavoprotein			2606135446	-	
Oxidative stress					
Catalase			2606134837	2606141744	
Peroxiredoxin			2606133036	2606138709	
			2606133255	2606139038	
			2606133445	2606139834	
			2606134579	2606141197	
			2606134983	2606142817	

			2606135825	2606142966	
			2606135851	2606143161	
			2606136211	2606143828	
			2606137559		
			2606137647		
Superoxide dismutase			2606136079	2606141687	
			2606136904		

Table S4.2. Carbohydrate-active enzymes encoded on the *Obscuribacterales* genomes

HMM	P3DObs1	IMG annotation	P3DObs2	IMG annotation
AA2	1	Peroxidase	0	
AA3	1	Choline dehydrogenase or related flavoprotein	0	
AA4	1	FAD/FMN-containing dehydrogenase	1	FAD/FMN-containing dehydrogenase
AA6	1	NAD(P)H-dependent FMN reductase	2	NAD(P)H-dependent FMN reductase
AA7	0		2	FAD/FMN-containing dehydrogenase
CBM16	2	Hypothetical protein	1	HEAT repeats/Protein of unknown function (DUF642)
CBM32	1	Beta-glucanase, GH16 family	0	
CBM34	1	Alpha amylase, catalytic domain/Alpha amylase, N-terminal ig-like domain	1	Glycosidase
CBM37	1	Beta-glucanase, GH16 family	0	
CBM40	1	Concanavalin A-like lectin/glucanases superfamily/Chitobiase/beta-hexosaminidase C-terminal domain	0	
CBM47	1	Uncharacterized protein, contains caspase domain	0	
CBM48	4	Malto-oligosyltrehalose trehalohydrolase; Isoamylase; glycogen branching enzyme (EC 2.4.1.18); Glycogen debranching enzyme GlgX	4	Malto-oligosyltrehalose trehalohydrolase; Isoamylase; Glycogen branching enzyme
CBM50	19	LysM domain; LysM domain/Peptidase_C39 like family; Uncharacterized conserved protein RhaS, contains 28 RHS repeats; Transposase and inactivated derivatives	20	LysM domain; RHS repeat-associated core domain
CBM56	2	Concanavalin A-like lectin/glucanases superfamily/Chitobiase/beta-hexosaminidase C-terminal domain	0	
CBM66	1	Choline dehydrogenase or related flavoprotein	0	
CE1	19	Lysophospholipase, alpha-beta hydrolase superfamily; Poly(3-hydroxybutyrate) depolymerase; Esterase/lipase superfamily enzyme; Predicted hydrolase of the alpha/beta-hydrolase fold; Prolyl oligopeptidase PreP, S9A serine peptidase family; Pimeloyl-ACP methyl ester carboxylesterase; Glutamyl peptidase. Serine peptidase. MEROPS family S09D	26	Poly(3-hydroxybutyrate) depolymerase; Lysophospholipase; Alpha-beta hydrolase superfamily; Dienelactone hydrolase; Pimeloyl-ACP methyl ester carboxylesterase; Prolyl oligopeptidase (EC:3.4.21.26). Serine peptidase. MEROPS family S09A; Predicted hydrolase of the alpha/beta-hydrolase fold; Esterase/lipase superfamily enzyme; Alpha/beta hydrolase of unknown function (DUF900); Glutamyl peptidase. Serine peptidase. MEROPS family S09D; Platelet-activating factor acetylhydrolase, isoform II
CE10	23	Lysophospholipase, alpha-beta hydrolase superfamily; Dipeptidyl	22	Lysophospholipase; Alpha-beta[202] hydrolase superfamily; Alpha/beta

		aminopeptidase/acylaminoacyl peptidase; Acetyl esterase/lipase; Fermentation-respiration switch protein FrsA, has esterase activity, DUF1100 family		hydrolase fold; Prolyl oligopeptidase family; Acetyl esterase/lipase; Dipeptidyl aminopeptidase/acylaminoacyl peptidase; Fermentation-respiration switch protein FrsA, has esterase activity, DUF1100 family; Esterase/lipase
CE11	1	UDP-3-O-acyl N-acetylglycosamine deacetylase	1	UDP-3-O-acyl-N-acetylglucosamine deacetylase
CE14	4	N-acetylglucosaminyl deacetylase, LmbE family	3	N-acetylglucosaminyl deacetylase, LmbE family
CE2	1	Phosphatidylserine/phosphatidylglycerophosphate/cardiolipin synthase or related enzyme	1	Lysophospholipase L1 or related esterase
CE3	4	Lysophospholipase L1 or related esterase; GDSL-like Lipase/Acylhydrolase family	2	Lysophospholipase L1 or related esterase
CE4	2	Peptidoglycan/xylan/chitin deacetylase, PgdA/CDA1 family; Polysaccharide deacetylase	2	Polysaccharide deacetylase; Peptidoglycan/xylan/chitin deacetylase, PgdA/CDA1 family
CE5	2	Alpha/beta hydrolase family; PE-PPE domain	0	
CE6	1	Domain of unknown function (DUF303)	2	Domain of unknown function (DUF303)
CE7	3	Alpha/beta hydrolase family; Lysophospholipase, alpha-beta hydrolase superfamily	3	Lysophospholipase, alpha-beta hydrolase superfamily
CE9	1	Dihydroorotase, multifunctional complex type	2	N-acetylglucosamine 6-phosphate deacetylase (EC 3.5.1.25); Dihydroorotase, multifunctional complex type
GH1	**4**	Beta-galactosidase; Broad-specificity cellobiase (EC 3.2.1.21); Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase	**3**	Broad-specificity cellobiase (EC 3.2.1.21); Beta-galactosidase
GH108	0		1	Predicted Peptidoglycan domain/Glycosyl hydrolase 108
GH109	8	Glyoxylase or a related metal-dependent hydrolase, beta-lactamase superfamily II; Predicted dehydrogenase	10	Predicted dehydrogenase; myo-inositol 2-dehydrogenase (EC 1.1.1.18); Glyoxylase or a related metal-dependent hydrolase, beta-lactamase superfamily II; Glycosyltransferase involved in cell wall biosynthesis;
GH110	1	Right handed beta helix region	0	
GH116	0		1	Glycogen debranching enzyme (alpha-1,6-glucosidase)
GH125	2	Meiotically up-regulated gene 157 (Mug157) protein (function unknown)	0	
GH13	11	Malto-oligosyltrehalose trehalohydrolase/ malto-oligosyltrehalose synthase; Isoamylase; Glycogen branching enzyme (EC 2.4.1.18); Trehalose synthase; Alpha-1,4-glucan:maltose-1-phosphate maltosyltransferase; Glycogen debranching enzyme GlgX; Alpha amylase, catalytic domain/Alpha	10	Malto-oligosyltrehalose trehalohydrolase; Malto-oligosyltrehalose synthase; Isoamylase; Glycogen branching enzyme (EC 2.4.1.18); Trehalose synthase; Alpha-1,4-glucan:maltose-1-phosphate maltosyltransferase; Glycosidase; Glycogen debranching enzyme GlgX; Isoamylase; Glycosidase

		amylase, N-terminal ig-like domain; Glycosidase; Pullulanase/glycogen debranching enzyme		
GH130	3	Predicted glycosyl hydrolase, GH43/DUF377 family	2	Predicted glycosyl hydrolase, GH43/DUF377 family
GH14	3	Glycosyl hydrolase family 14	2	Glycosyl hydrolase family 14
GH15	2	Glucoamylase (EC:3.2.1.3)	1	Glucoamylase (EC:3.2.1.3)
GH16	1	Beta-glucanase, GH16 family	0	
GH18	2	Putative hydrolase, CocE/NonD family; Protein of unknown function (DUF3142)	1	Glycosyl hydrolases family 18
GH23	5	Transglycosylase SLT domain; Soluble lytic murein transglycosylase and related regulatory proteins (some contain LysM/invasin domains)	4	Transglycosylase SLT domain
GH25	2	Lyzozyme M1 (1,4-beta-N-acetylmuramidase), GH25 family	5	Lyzozyme M1 (1,4-beta-N-acetylmuramidase), GH25 family;
GH26	**0**		**1**	Glycosyl hydrolase family 26
GH28	0		1	Glycosyl hydrolases family 28
GH3	**4**	Periplasmic beta-glucosidase and related glycosidases	**3**	Periplasmic beta-glucosidase and related glycosidases
GH30	1	O-Glycosyl hydrolase	2	O-Glycosyl hydrolase
GH31	1	Alpha-glucosidase, glycosyl hydrolase family GH31	2	Alpha-glucosidase, glycosyl hydrolase family GH31
GH33	0		1	BNR repeat-like domain
GH35	**4**	Glycosyl hydrolases family 35	**2**	Glycosyl hydrolases family 35
GH38	**1**	Alpha-mannosidase	**1**	Alpha-mannosidase
GH46	0		1	Hypothetical protein
GH5	**2**	Sugar-binding cellulase-like; Aryl-phospho-beta-D-glucosidase BglC, GH1 family	**3**	Cellulase (glycosyl hydrolase family 5); Aryl-phospho-beta-D-glucosidase BglC, GH1 family
GH55	1	Hypothetical protein	**0**	
GH57	1	Predicted glycosyl hydrolase, contains GH57 and DUF1957 domains	1	Predicted glycosyl hydrolase, contains GH57 and DUF1957 domains
GH65	0		1	Haloacid dehalogenase superfamily, subfamily IA, variant 3 with third motif having DD or ED/beta-phosphoglucomutase family hydrolase
GH73	1	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase	1	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase
GH75	2	Fungal chitosanase of glycosyl hydrolase group 75	2	Fungal chitosanase of glycosyl hydrolase group 75
GH76	3	Glycosyl hydrolase family 76; Uncharacterized conserved protein YyaL, SSP411 family, contains thoiredoxin and six-hairpin glycosidase-like domains	2	Glycosyl hydrolase family 76; Uncharacterized conserved protein YyaL, SSP411 family, contains thoiredoxin and six-hairpin glycosidase-like domains
GH77	3	Malto-oligosyltrehalose synthase; 4-alpha-glucanotransferase	3	Malto-oligosyltrehalose synthase; 4-alpha-glucanotransferase
GH78	3	Glycogen debranching enzyme (alpha-1,6-glucosidase); Glycogen debranching enzyme, archaeal type, putative	2	Glycogen debranching enzyme (alpha-1,6-glucosidase); Glycogen debranching enzyme, archaeal type, putative
GH93	1	BNR repeat-like domain	0	

GT1	0		1	Uncharacterized membrane protein, YccA/Bax inhibitor family
GT19	1	Lipid-A-disaccharide synthase (EC 2.4.1.182)	1	Lipid-A-disaccharide synthase (EC 2.4.1.182)
GT2	33	Glycosyltransferase involved in cell wall biosynthesis; Glycosyltransferase, catalytic subunit of cellulose synthase and poly-beta-1,6-N-acetylglucosamine synthase; Glycosyltransferase, GT2 family	40	Glycosyltransferase involved in cell wall biosynthesis; Glycosyl transferase family 2; Glycosyltransferase, catalytic subunit of cellulose synthase and poly-beta-1,6-N-acetylglucosamine synthase; Nucleoside-diphosphate-sugar epimerase; Methyltransferase, FkbM family
GT20	3	Trehalose 6-phosphate synthase (EC 2.4.1.15)	2	Trehalose 6-phosphate synthase (EC 2.4.1.15)
GT21	1	Glycosyltransferase, catalytic subunit of cellulose synthase and poly-beta-1,6-N-acetylglucosamine synthase	1	Glycosyltransferase, catalytic subunit of cellulose synthase and poly-beta-1,6-N-acetylglucosamine synthase
GT26	1	N-acetylmannosaminyltransferase (EC 2.4.1.187)	1	N-acetylmannosaminyltransferase (EC 2.4.1.187)
GT27	2	Glycosyltransferase, GT2 family	2	Glycosyltransferase involved in cell wall biosynthesis; Glycosyltransferase, GT2 family
GT28	3	UDP-N-acetylglucosamine:LPS N-acetylglucosamine transferase	3	Monogalactosyldiacylglycerol (MGDG) synthase/Glycosyltransferase family 28 C-terminal domain; UDP-N-acetylglucosamine-N-acetylmuramylpentapeptide N-acetylglucosamine transferase
GT30	1	3-deoxy-D-manno-octulosonic-acid transferase	1	3-deoxy-D-manno-octulosonic-acid transferase
GT35	2	Glycogen/starch/alpha-glucan phosphorylases	2	Glycogen/starch/alpha-glucan phosphorylases
GT39	3	Dolichyl-phosphate-mannose-protein mannosyltransferase; hypothetical protein	3	Dolichyl-phosphate-mannose-protein mannosyltransferase
GT4	24	Glycosyltransferase involved in cell wall biosynthesis; Glycosyl transferases group 1; N-acetyl-alpha-D-glucosaminyl L-malate synthase BshA; Glycosyl transferases group 1/Glycosyl transferase 4-like domain; Hypothetical protein	20	Glycosyltransferase involved in cell wall biosynthesis; N-acetyl-alpha-D-glucosaminyl L-malate synthase BshA; Glycosyl transferases group 1; Glycosyl transferases group 1/Glycosyl transferase 4-like domain
GT45	2	Signal recognition particle receptor subunit beta, a GTPase; ADP-ribosylation factor family	1	Signal recognition particle receptor subunit beta, a GTPase
GT51	5	Membrane carboxypeptidase (penicillin-binding protein); Penicillin-binding protein, 1A family; Transglycosylase	4	Penicillin-binding protein, 1A family
GT8	0		2	Lipopolysaccharide biosynthesis protein, LPS:glycosyltransferase
GT83	6	Dolichyl-phosphate-mannose-protein mannosyltransferase; Uncharacterized membrane protein	10	Dolichyl-phosphate-mannose-protein mannosyltransferase; 4-amino-4-deoxy-L-arabinose transferase or related glycosyltransferase of PMT family
GT87	5	Protein of unknown function (DUF2029)	5	Protein of unknown function (DUF2029)

GT9	4	ADP-heptose:LPS heptosyltransferase; Glycosyltransferase family 9 (heptosyltransferase)	1	ADP-heptose:LPS heptosyltransferase
GT94	0		1	Glycosyltransferase involved in cell wall bisynthesis
PL12	1	Heparinase II/III-like protein	1	Heparinase II/III-like protein
PL22	1	Dipeptidyl aminopeptidase/acylaminoacyl peptidase	2	Uncharacterized N-terminal domain of tricorn protease; WD40-like Beta Propeller Repeat
SLH	2	S-layer homology domain	4	S-layer homology domain
Total	270 (4.1%)		272 (4.4%)	

Table S4.3. antiSMASH results for the *Obscuribacterales* genomes

antiSMASH result	IMG accession #	IMG annotation
P3DObs1		
Bacteriocin	2606132669	Uncharacterized conserved protein, UPF0276 family
Bacteriocin	2606135539	Uncharacterized conserved protein, UPF0276 family
Bacteriocin	2606135789	Protein of unknown function (DUF692)
Bacteriocin	2606136253	Uncharacterized conserved protein, UPF0276 family
TypeIIIpks	2606132052	Predicted naringenin-chalcone synthase
Other KS	2606132250	Acyl transferase domain in polyketide synthase (PKS) enzymes
Other KS	2606132251	Acyl transferase domain in polyketide synthase (PKS) enzymes
P3DObs2		
Bacteriocin	2606140554	Protein of unknown function (DUF692)
Bacteriocin	2606143737	Uncharacterized conserved protein, UPF0276 family
Bacteriocin	2606143979	Uncharacterized conserved protein, UPF0276 family
Hglks	2606139360	Acyl transferase domain in polyketide synthase (PKS) enzymes
Hglks	2606139361	Acyl transferase domain in polyketide synthase (PKS) enzymes
TypeIIIpks	2606143013	Predicted naringenin-chalcone synthase

Appendix D: Screening and visualising Melainabacteria

Screening for Melainabacteria 16S rRNA genes

A range of methods targeting the 16S rRNA gene was developed to identify members of the Melainabacteria using PCR, real-time PCR and amplicon sequencing. Methods for visualising members of the Melainabacteria were also used including sub-optimal multi-probe fluorescence *in situ* hybridisation (SUMP-FISH) with Transmission Electron Microscopy (TEM) and FISH with helper oligonucleotide probes.

Small-subunit (SSU) amplicon sequencing

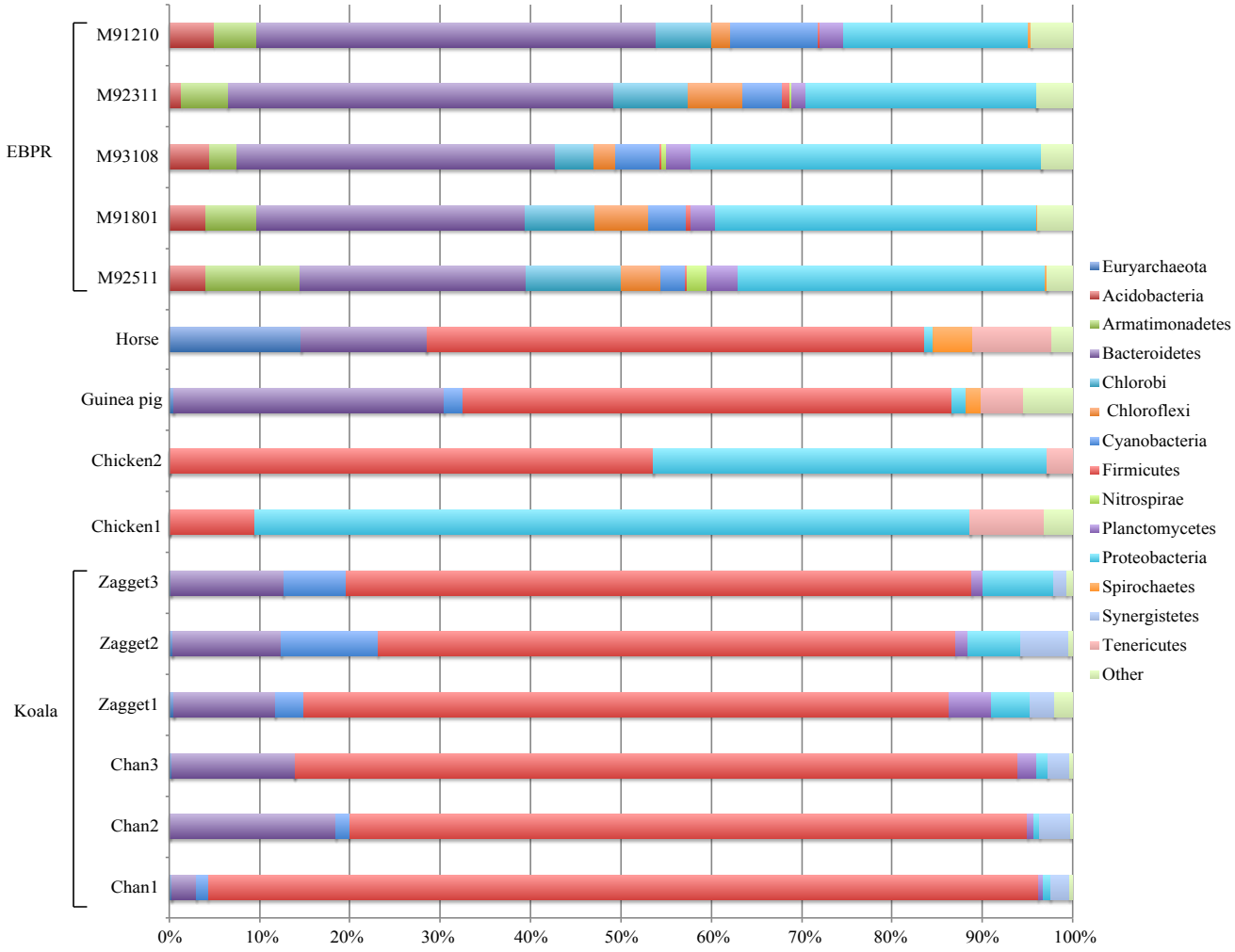
Recently a plethora of high-throughput DNA sequencing technologies has become available, such as the 454 GS FLX and GS Junior (Roche), MiSeq and HiSeq2000 (Illumina), SOLiD (Applied Biosystems/Life-Technologies), IonTorrent PGM (Life Technologies) and the PacBio RS from Pacific Biosciences [1-3]. These sequencing platforms provide large amounts of data, are less time consuming and are more cost effective than traditional Sanger sequencing. A large number of samples can be run in parallel by multiplexing and samples can be split based on unique sample-specific barcodes, in a process called SSU amplicon sequencing [4].

Methods

Genomic DNA (gDNA) was extracted from horse, chicken, koala and guinea pig faeces using a PowerSoil DNA Isolation Kit (MoBio, CA, USA) as per instructions but with two additional ethanol washes. The gDNA from an enhanced biological phosphorous removal (EBPR) reactor was extracted using a FastDNA Spin Kit (MP Biomedicals, CA, USA) as per instructions. The V6 to V8 regions of the 16S rRNA gene was amplified using fusion primers containing 454 adaptor sequences ligated to the primers 926F (5'- AAACTYAAAKGAATTGRCGG-3') and 1392R (5'-ACGGGCGGTGTGTRC-3') [5]. Multiplex identifiers consisting of five nucleotides were incorporated in the 1392R primer to allow for multiplexing. Fifty microlitre PCR reactions were prepared containing 20 ng of template DNA, 5 µL of 10x buffer (Fisher Biotec, Wembley, Australia), 1 µL of 10 mM dNTP mix (Fisher Biotec) 1.5 µL BSAI (Fisher Biotech), 4 µL 25 mM $MgCl_2$ (Fisher Biotec), 1 µL of each 10 µM primer, and 1 unit of *Taq* polymerase (Fisher Biotec). Cycling conditions were 95°C for 3 mins, followed by 30 cycles of 95°C for 30 s, 55°C for 30 s and 74°C for 30 s followed by a final extension of 74°C for 10 mins. Following amplification, PCR products for each sample were purified using the Agencourt AMPure XP PCR purification system

(Beckman-Coultier, CA, USA) and quantified using the Qubit Fluorometer (Invitrogen, CA, USA). Amplicons were sequenced from the reverse primers using the Roche 454 GS-FLX Titanium platform (Roche, CT, USA) at the Australian Centre for Ecogenomics (University of Queensland, Australia) (ACE, UQ, Australia). Sequence data generated were demultiplexed and processed using a modified version of the QIIME pipeline [6], which uses Acacia 1.50 (app v 2.0.0) [7] to correct homopolymer errors (modified pipeline is available at https://github.com/Ecogenomics/APP). Sequences were clustered at 97% sequence identity and the taxonomy of the representatives from each operational taxon unit (OTU) was assigned using BLASTN v. 2.2.26 [8] against the Greengenes database, version 12_10) [9].

Figure 1. Relative phylum abundances for sequence reads in each sample



All EBPR samples from M9 contained Melainabacteria cells in the order Obscuribacterales and all koala and guinea pig samples contained Melainabacteria cells in the order Gastranaerophilales. The horse, chicken1 and chicken2 samples did not contain any Melainabacteria cells.

Conclusion

The EBPR samples from M9 was used as a positive control for primers and probes designed to target members from the order Obscuribacterales and the koala and guinea pig samples were used as a positive control for primers and probes used to target members from the order Gastranaerophilales. The horse and chicken samples were used as a negative control.

Polymerase chain reaction (PCR)

The use of the 16S rRNA as a marker gene revolutionised molecular phylogenetics resulting in the identification of bacteria that had not previously been seen before with cultivation methods. PCR primers were designed as a way to screen multiple samples for Melainabacteria in a range of environments.

Methods

Primer design

The Greengenes 13_05 database [10] was loaded into ARB [11] and the PROBE_DESIGN option was used to design probes that would target conserved regions of the 16S rRNA gene at both the class and order-level of Melainabacteria.

Table 1. Probes designed for PCR

Probe	**E.coli no.**	**Probe sequence**	**Target sequence**	**Target group**
DarCy492F	492-509	CAGARGAAGCATCGGCTA	CAGARGAAGCAUCGGCUA	Class Melainabacteria
DarCy304F	304-321	TGATCAGCCACAATGGGA	UGAUCAGCCACAAUGGGA	Order Gastranaerophilales
Obs874F	874-891	GTATCCCGCCTGAGUAGT	GUAUCCCGCCUGAGUAGU	Order Obscuribacterales

Forward primer	**Reverse primer**	**Size of product (bp)**
DarCy492F	907R [12]	415
DarCy304F	907R [12]	603
Obs874F	1392R [13]	518

Different optomisation conditions used

1. Temperature gradient (50-65°C)
2. $MgCl_2$ gradient (0.5mM-4mM)
3. Cycle number (32 and 35)
4. *Taq* concentration (0.1µl, 0.2µl at 5U/µL)
5. Primer concentration (2.5µM and 5µM)

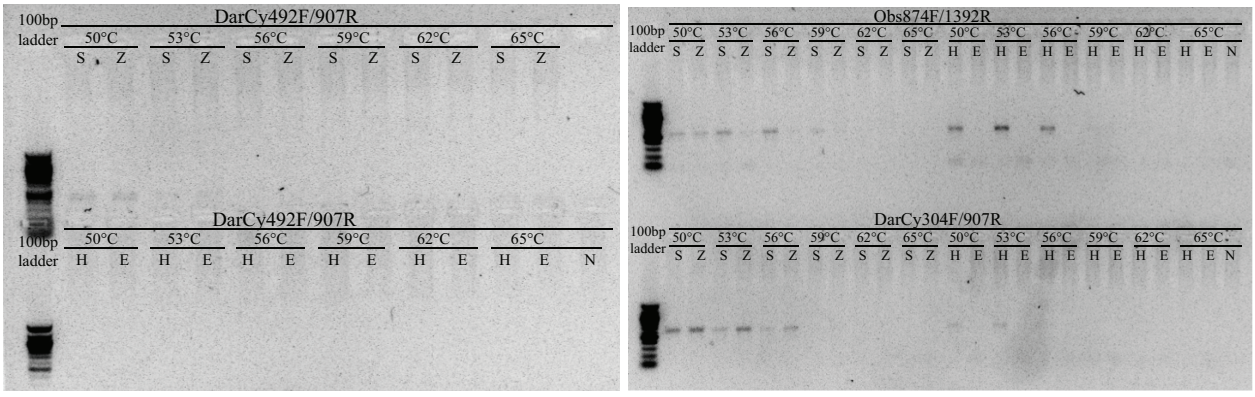
Samples used

1. Koala faeces (+ve for Melainabacteria, order Gastranaerophilales)
2. Enhanced biological phosphorous removal reactor (Sequencing batch reactor) (SBR) (+ve for Melainabacteria, order Obscuribacterales)
3. *Escherichia coli* (–ve)
4. Horse faeces (–ve)
5. Blank

Fifty microliters PCR reactions were prepared containing 20 ng of template DNA, 5 µL of 10x buffer (Fisher Biotech), 1 µL of 10 mM each dNTP mix (Fisher Biotec), 0.5 mM-4 mM $MgCl_2$ (Fisher Biotec), 0.1 µl-0.2 µl of 5U/µL taq (Fisher Biotec), 0.25-0.5 µL of each 10 µM primer (Fisher Biotec), 1.5 ul of BSA (Fisher Biotec) and the remaining was MQ water. The PCR was run with cycling conditions of 95°C for 10 mins, followed by 32 or 35 cycles of 95°C for 3 mins, 95°C for 30 s, 50-65°C for 30 s and 72°C for 30 s followed by 95°C for 2 mins and 60°C for 15 s.

Results

Figure 2. PCR results using primers designed to target Melainabacteria 16S rRNA genes



Primers are labelled above the lanes. Samples are indicated by abbreviations: Sequencing batch reactor (S), koala (Zagget) (Z), horse (H), *E. coli* (E) and negative (N).

The DarCy492F primer was designed to target all Melainabacteria and was successful at amplifying the 16S rRNA genes from SBR and koala faeces at the lowest temperature tested (50°C). It was also

negative for horse faeces and *E. coli*. The Obs874F primer designed to target members of the order Obscuribacterales successfully amplified the SBR sample. However, the negative samples (koala and horse faeces) were also amplified. By increasing the temperature to 59°C, there was no amplification of the negative samples but only a weak amplification of the SBR sample. The DarCy304F primer, designed to target members of the order Gastranaerophilales successfully amplified koala faeces but there was also amplification of the negative samples, SBR and horse faeces.

Conclusion

There was some success using the DarCy492F primer for amplification of all Melaianbacteria. However in later experiments the *E. coli* (negative control) was also amplified. The primers, Obs874F and DarCy304F were non-specific for orders Obscuribacterales and Gastranaerophilales.

Real-time PCR

Real-time PCR can be used to measure the DNA concentration continuously during amplification. This enables the initial template concentration to be determined and cell numbers to be accurately counted [14]. This method is more accurate than end point detection PCR and 16S rRNA gene-targeted group-specific primers can be designed to target a group of interest [5].

Methods

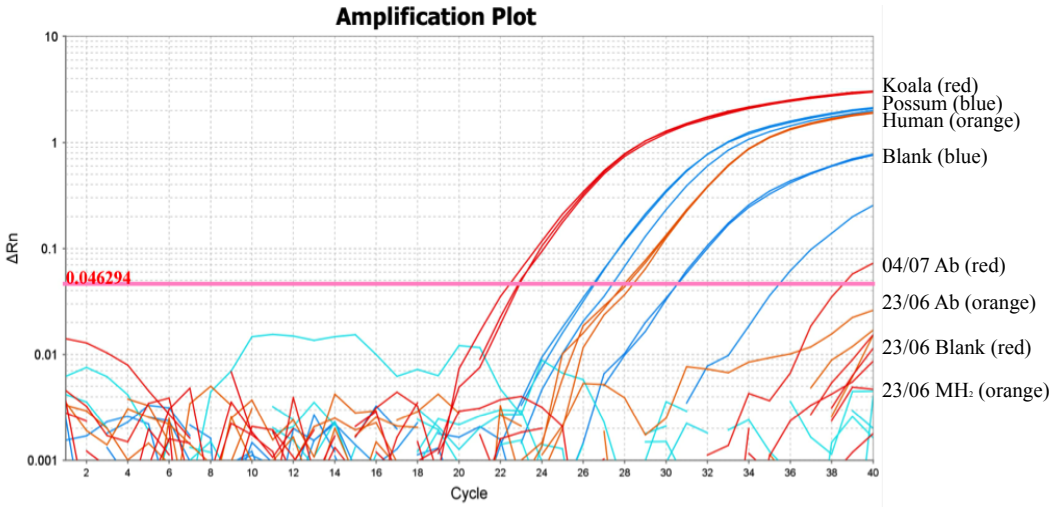
DNA extraction, primer design and real-time PCR

Genomic DNA was extracted from koala, possum and human faeces using a PowerSoil DNA Isolation Kit (MoBio, Carlsbad, CA, USA) as per instructions but with an additional two ethanol washes. The Greengenes 13_05 database [10] was loaded into ARB [11] and the PROBE_DESIGN option was used to design probes that would target the 16S rRNA gene from the order Gastranaerophilales. The V7-V8 regions of the 16S rRNA gene was amplified using Gast1261F (5'-TGGAACAGCGAGCAGCGA-3')/Gphas1151 (5'-CTGTGGCTATGGCTCTCT-3') and 1392R (5'-ACGGGCGGTGTGTRC-3') [5]. One hundred microlitre PCR reactions were prepared containing different dilutions of DNA, 1 μL of each 10 μM primer and 5 μl of 2x SYBR Green/AmpliTaq Gold DNA polymerase mix (Life Technologies, CA, USA). Primers targeting the *E. coli* rpsL gene (forward primer: 5′-GTAAAGTATGCCGTGTTCGT-3′, and reverse primer: 5′-AGCCTGCTTACGGTCTTTA-3′) [15] was used as an inhibition control, amplifying *E. coli* DH10B genome DNA only. Three dilutions 1/100, 1/500 and 1/1000 (microbial template DNA, Gast1261F/1392wR) were run in triplicate for each sample. The PCR was run on the ViiA7 Real-

Time PCR System (Life Technologies) with cycling conditions of 95°C for 10 mins, followed by 40 cycles of 95°C for 15 s, 52°C for 20 s and 72°C for 30 s followed by 95°C for 2 mins and 60°C for 15 s. A melt curve was produced by running a cycle of 2 mins at 95°C and a last cycle of 15 s at 60°C. The cycle threshold (Ct) values were recorded and analysed using ViiA7 v1.2 software (Life Technologies).

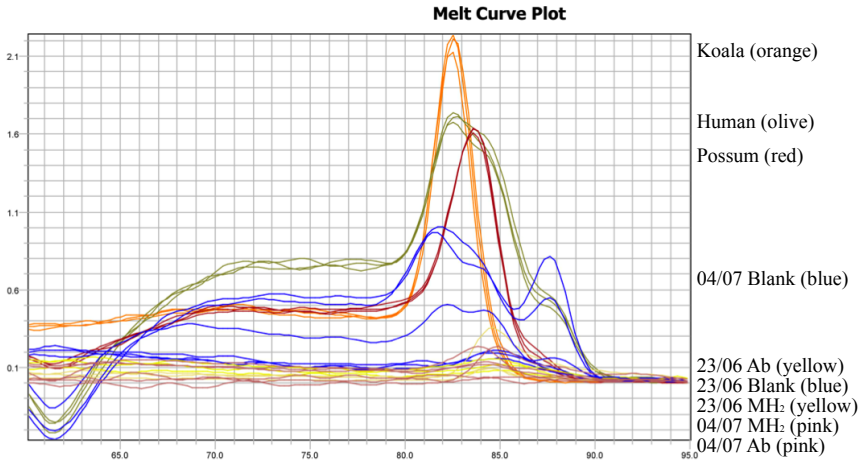
Results

Figure 3. Amplification plot of real-time PCR for Melainabacteria



Amplicon sequencing data indicated that the possum faeces contained the highest abundance of Gastranaerophilales, followed by the koala faeces. However, the koala faeces amplified in an earlier cycle and the possum faeces had a similar profile to human faeces, which was negative for Gastranaerophilales.

Figure 4. Melt curve plot of the real-time PCR for Melainabacteria



The melt curve results showed that koala faeces had the highest amplification, whereas the possum and human faeces had a similar profile.

Conclusion

It would be expected that the possum faeces would amplify first and show the highest peak as it was indicated by amplicon sequencing data that it contains the highest abundance of Gastranaerophilales. However, the primers may be more specific towards the populations found in the koala faeces as the primers had been designed to target these populations.

Visualising Melainabacteria using the 16S rRNA gene as a target with Fluorescent *in situ* hybridisation (FISH)

Fluorescent *in situ* hybridisation (FISH) works by detecting nucleic acid sequences with a fluorescently labelled probe that hybridises specifically to its complementary target sequence with the intact cell. The most common target for FISH is the SSU rRNA gene because of its genetic stability, it has both conserved and variable regions and it is naturally amplified in the cell [16].

FISH is a technique that can provide information about phylogenetic identity, number, morphology and spatial distribution of microorganisms without the need for cultivation [17]. FISH can be used as a way to detect organisms of interest, to help in understanding complex microbial communities and to a lesser degree to describe the physiological state of a cell. FISH can also be used to look at organisms and how they interact with other organisms.

Sub-optimal multi-probe fluorescent *in situ* hybridisation (SUMP-FISH), flow sorting and Transmission Electron Microscopy (TEM)

SUMP-FISH was designed to target multiple orders within the class Melainabacteria. The idea was to use multiple FISH probes that would attach to different regions of the 16S rRNA gene. It was hoped that by using multiple probes there would be more fluorescence than a single probe. The samples would then be sorted using flow cytometry and TEM would be performed to visualise the cells.

Methods

SUMP-FISH

Probes for SUMP-FISH were designed using the design probe module in ARB with the target group 180 Melainabacteria in the Greengenes database [10]. The settings for the SUMP-FISH probe design was 75% target group with < 5,000 non-target group hits. Seven probes were chosen as in

Table 2 with a Cy3 fluorophore attached to the 3' end. Multiple probes in close proximity were chosen to assist in opening the hairpin turns in the 16S rRNA gene.

Table 2. Probes designed for SUMP-FISH

Probe	**E.coli no.**	**Probe sequence**	**Target sequence**
DarCy312	312-329	CACAATGGGACTGAGACA	TGTCTCAGTCCCATTGTG
DarCy330	330-347	CCGTAGGAGTATGGGCCG	CGGCCCATACTCCTACGG
DarCy360	360-377	CCATTGCGCAAAATTCCC	GGGAATTTTGCGCAATGG
DarCy1219	1219-1236	TAACACGTGTGTAGCCCA	TGGGCTACACACGTGTTA
DarCy1336	1336-1352	GTTTACTAGCGATTCCG	CGGAATCGCTAGTAAAC
DarCy1353	1353-1369	GCAGCGTGCTGATCTGC	GCAGATCAGCACGCTGC
DarCy1370	1370-1387	CCCGGGAACGTATTCAAC	GTTGAATACGTTCCCGGG

Fixation-free FISH was used for enrichment of microbial populations without the use of paraformaldehyde [18]. Koala faeces from Zagget the koala and EBPR samples were used for SUMP-FISH. Samples were pelleted by centrifugation and washed twice with phosphate buffered saline (PBS) before being stored in 10% glycerol at -20°C.

Fifty microlitres of prepared fixed cells were washed twice with 500 µl PBS and centrifuged at 10,000g for 3 mins. Samples were washed with 500 µl of 50%, 80% and 100% ethanol and placed in a heating block at 4°C to dry. Thirty percent hybridisation buffer was made from 360 µl of 5M NaCl, 40 µl Tris, 600 µl formamide, 1 ml of MQ and 2 µl of 10% SDS. Forty five microliters of hybridisation buffer was added to each pellet, which was resuspended. Five microliters of each of the probes at 25 ng/µl was added to each sample and 5ul of buffer was added to the control (no probe). The samples were placed in aluminium foil and hybridized at 46°C for 2.5 hours. Washing buffer was made by adding 112 µl of 5M NaCl, 1 ml of Tris/HCl, 500 µl of 0.5M EDTA, 50 µl of 10% SDS and up to 50 ml of MQ. Five hundred microliters of 50°C washing buffer was added to each tube. The samples were centrifuged at 10,000 rpm for 3 mins, the supernatant was removed and another 500 µl of washing buffer was added. The samples were vortexed and placed in a 48°C waterbath for 10-15 mins. The samples were then centrifuged for 3 mins at 10,000 rpm and the supernatant was removed. The pellets were washed with 500 µl of ice cold PBS, vortexed and

centrifuged for 3 mins at 10,000 rpm. The supernatant was removed and the pellet was resuspended in 500 µl of cold PBS. Samples were sonicated in a waterbath for 30 secs. Samples were centrifuged for 3 mins at 10,000 rpm and resuspended in 500 µl of MQ water.

Flow sorting was performed with a BD FACS ARIA III (BD Biosciences, CA, USA) in two stages to reduce final sort volume. A 561-nm laser was used as the excitation source for light scattering and fluorescence. Homogenised unfixed samples were diluted to 10^6-10^7 cells ml^{-1} to prevent nozzle clogging and sorted first at high speed (25,000 $cells^{-1}$) into 50 ml centrifuge tubes on ice. Followed by resorting at a lower speed (10,000 cells s^{-1}) into an eppendorf tube.

PCR of sorted fixed cells

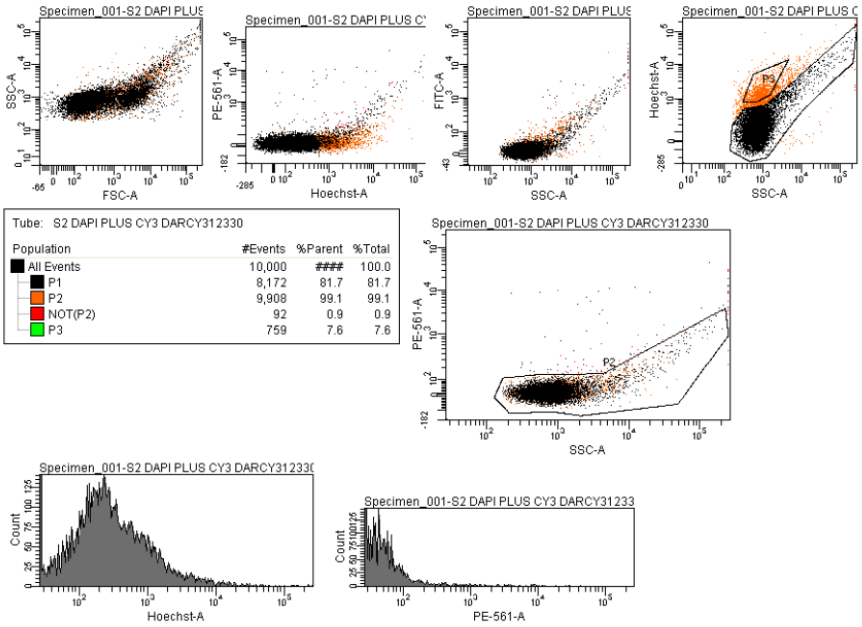
Twenty five thousand sorted cells were centrifuged at 10,000 rpm for 3 mins and resuspended in 10 µl of MQ water. Three cycles of freeze thaw was used to break open the cells for PCR. One to four microliters of sorted cells was used as input for PCR and the reaction was run as above with primers 803F (5'-TTAGAKACCCBNGTAGTC-3') and 1392R (5'-ACGGGCGGTGTGTRC-3') [5].

Transmission Electron Microscopy

Sapphire discs were carbon coated and then coated with poly L-lysine. The cells from the flow sorter were centrifuged for 3 mins at 10,000 rpm. The supernatant was removed, leaving 2 µl of liquid at the bottom of the tube. Cells were scraped from the sides and resuspended before placing the cells on a sapphire disc. The discs were dipped in 20x bovine serum albumin and the cells were left to settle on the disc for 9 mins. The discs were placed in a hat that had been coated in hexadicine. The sample was placed into a high-pressure freezer and samples were collected with liquid nitrogen into a cryovial which had holes punched in them to release the pressure. Samples were then stored at -80°C until further processing. Samples were cut into thin sections and viewed with a JEOL JEM2100 LaB_6 stem transmission electron microscope at the Centre for Microscopy and Microanalysis (UQ, Australia). A Gatan Orius 100 slow scan CCD camera was used to capture images.

Results

Figure 5. FACS ARIA results of sorted cells from SBR9



The black circles were identified as non-Melainabacteria cells and the orange circles were identified as potential Melainabacteria cells and were collected for TEM.

Figure 6. TEM images of the cells that were sorted using the SUMP-FISH probes

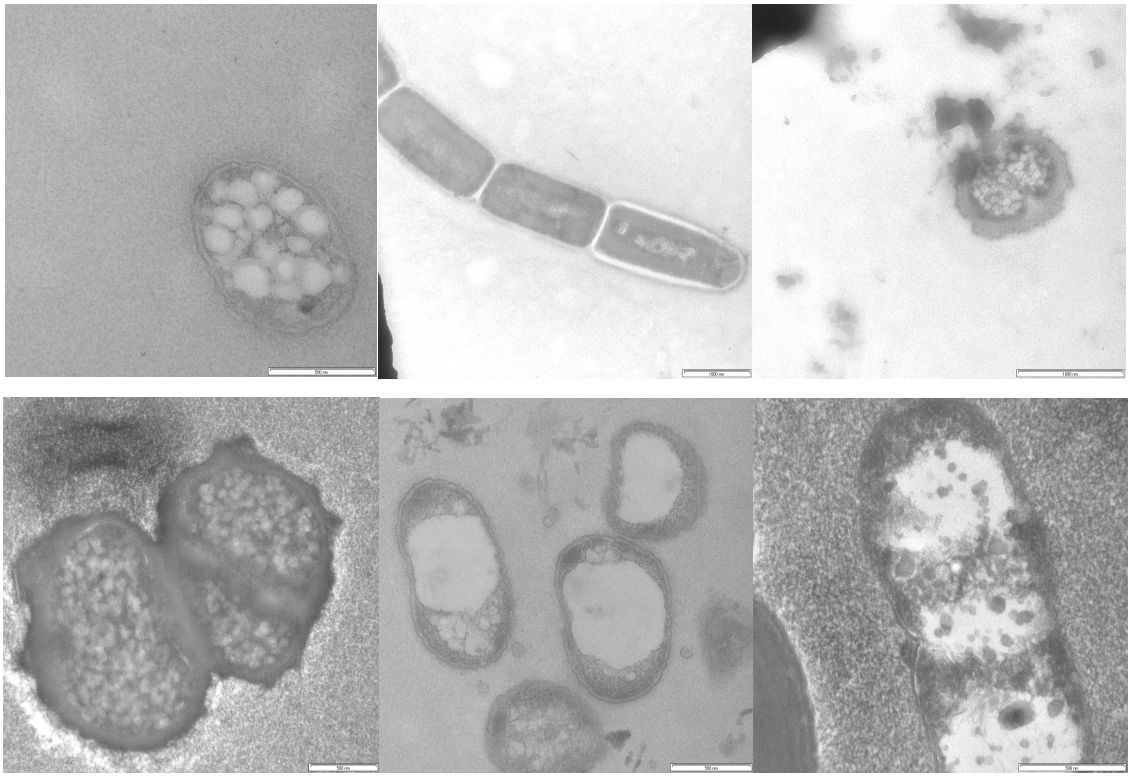
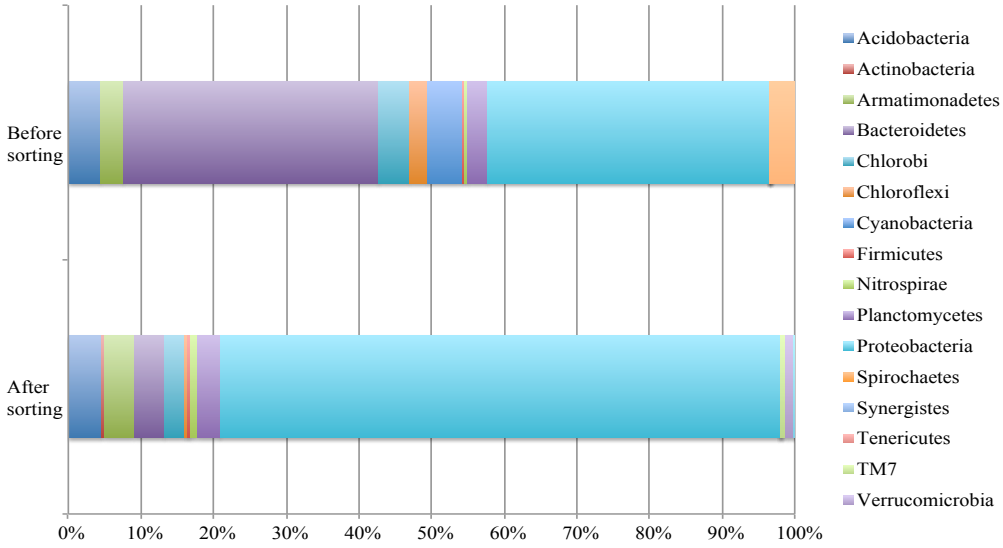


Figure 7. Bar plot of amplicon sequencing data before and after sorting for Melainabacteria



The amplicon sequencing results showed that no Melainabacteria representatives were sorted by the flow sorter.

Conclusion

We decided that SUMP-FISH was not the best method to isolate melainabacterial cells because it was not specific enough. Single FISH probes with helper oligonucleotide probes were then designed to be more specific to the 16S rRNA genes of the Melainabacteria.

Fluorescent *in situ* Hybridisation (FISH) with helper oligonucleotide probe

Methods

Samples were vortexed with paraformaldehyde and fixed for 2 hours. Samples were then centrifuged and the paraformaldehyde was removed. PBS was added to the samples, which were then centrifuged and the PBS was removed. Samples were stored in 1:1 Ethanol:PBS in a -20°C freezer until needed.

The EBPR sample washed off the slides during the washing stage so gelatin was added to the slides before adding the samples. Gelatin was prepared by adding 0.1% gelatin with 0.01% chromium potassium sulphate. Approximately 2 µl of sample and ~2 µl of gelatin was added to the slides and dried overnight in the dark. The slides were dried in an ethanol series (3 mins in each) of 50%, 80% and 98% ethanol. Following, slides were airdried. Thirty percent hybridisation buffer was prepared

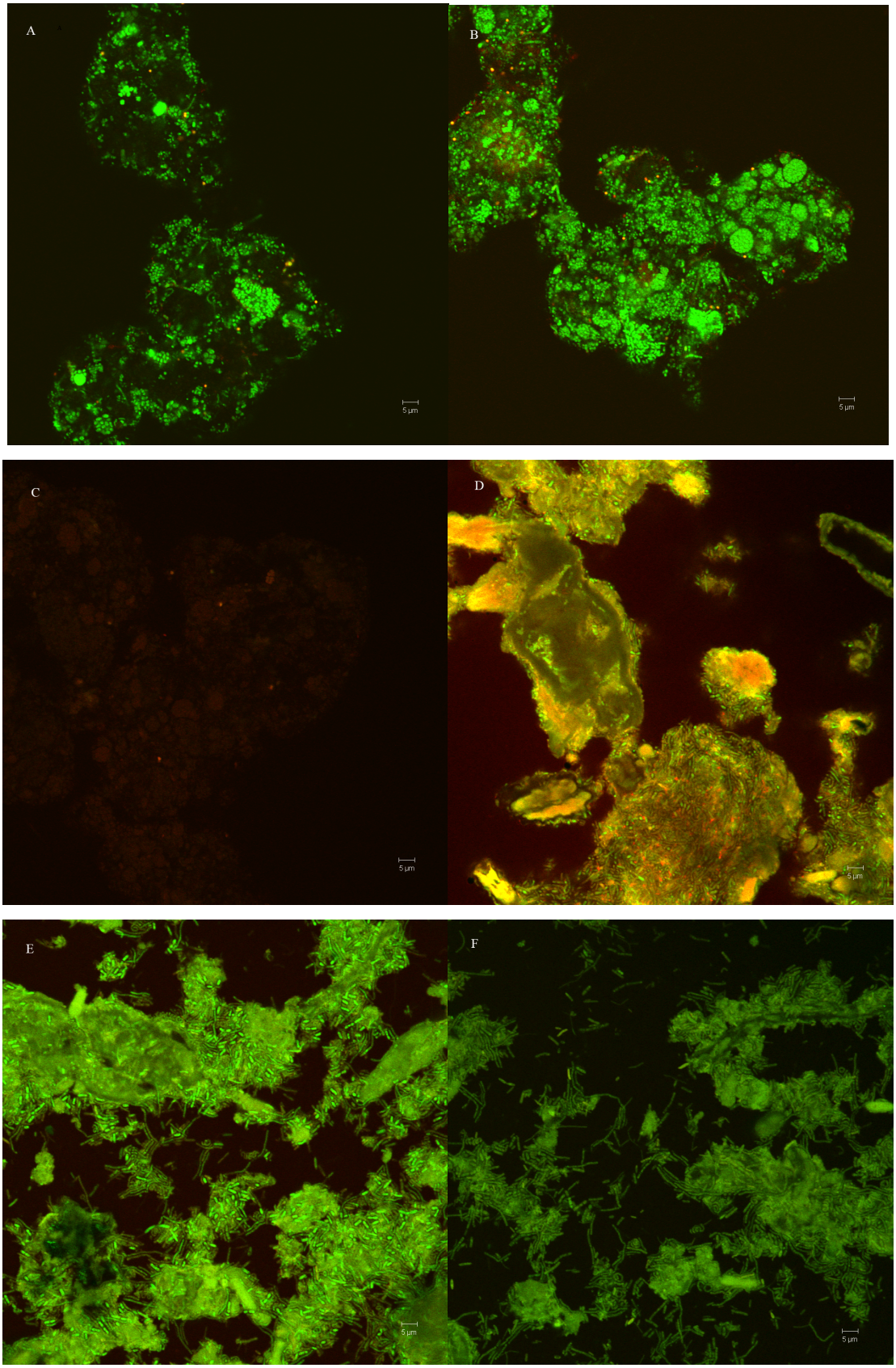
in a 2ml eppendorf tube and consisted of 360 µls of 5 M NaCl (autoclaved), 40 µl of 1M Tris/HCl (autoclaved), 600 µl of formamide, 998 µl of autoclaved MQ water and 2 µl of 10% SDS. Hybridisation buffer was added to each well on the slides and the remainder was used to moisten a kimwipe, which was placed in a 50 ml falcon tube. One microlitre of probe from the stock concentration of 25ng/µl was added and mixed carefully. The slides were placed in the 50 ml falcon tube containing the moistened kimwipe. The tubes were placed in a hybridisation oven at 46°C for 2 hours. Wash buffer was prepared in a 50 ml falcon tube and consisted of 1.02 mls of 5M NaCl, 1 ml of 1M Tris/HCl, 500 µl of 0.5M ethylenediaminetetraacetic acid (EDTA), up to 50 mls of MQ water and 50 µl of 10% SDS. After hybridisation, the slides were carefully removed from the falcon tube and rinsed immediately with wash buffer by pipetting a small amount of wash buffer over the slide. The slides were then placed into the tube containing the wash buffer and left to incubate in a water bath at 48°C for 10 mins. The slides were removed and gently immersed in a beaker containing ice-cold water for ~ 2-3 secs to remove the salt. The slide was vacuum dried. An anti-fade mounting media, 1,4-Diazabicyclo [2.2.2] octane (DABCO) (Sigma) was added to the slides and a coverslip was applied. The slides were viewed with a Zeiss LSM510 confocal microscope.

Table 3. Probes designed for FISH

Probe	E.coli no.	Probe sequence	Target sequence	Specificity
DarCy304	304-321	TCCCATTGTGGCTGATCA	UGAUCAGCCACAAUGGGA	most f__Gastranaerophilaceae
DarCy304H1	283-302	CCTCTCARACCARCTACYGA	UCRGUAGYUGGUYUGAGAGG	
DarCy304H2	323-342	GGAGTATGGGCCGTRTCTCA	UGAGAYACGGCCCAUACUCC	
DarCy492	492-509	TAGCCGATGCTTCYTCTG	CAGARGAAGCAUCGGCUA	most c__Melainabacteria and c__ML635J-21
DarCy492H1	447-490	GGGTACCGTCANNNTTCGTC	GACGAANNNUGACGGUACCC	
DarCy492H2	511-529	CGCGGCTGCTGGCACGKAG	CUMCGUGCCAGCAGCCGCG	
VVIB1147	1147-1162	GGCAGTCTGGCCTGAGGG	CCCUCAGGCCAGACUGCC	most of o__Vampirovibrionales
VVIB1147H1	1112-1129	GTAWCAACAGACMACRAGGG	CCCUYGUKGUCUGUUGWUAC	
VVIB1147H2	1164-1183	ACCTTCCTCCGGTTTRTCAC	GUGAYAAACCGGAGGAAGGU	
Obs874	874-891	ACTACTCAGGCGGGATAC	GUAUCCCGCCUGAGUAGU	o__Obscuribacteriales
Obs874H1	853-872	TABCGCGTTWRCTRCGGCAC	GUGCCGYAGYWAACGCGVUA	
Obs874H2	893-912	GAGTTTCAACCTTGCGGCCG	CGGCCGCAAGGUUGAAACUC	

H1 and H2 refer to helper 1 and helper 2 probes

Figure 8. FISH images of SBR and koala faeces



A. SBR using probes Obs874 (Cy3) and EUB338 (FITC) at 10x magnification. **B.** SBR using probes DarCy492 (Cy3) and EUB338 (FITC) at 10x magnification. **C.** SBR with no probes at 10x magnification. **D.** Koala using DarCy304 (Cy3) and EUB338 (FITC) at 40x magnification. **E.** Koala using Darcy492 (Cy3) and EUB338 (FITC) at 40x magnification. **F.** Koala with no probes at 40x magnification.

Issues with autofluorescence

The koala faecal samples contained plant material resulting in autofluorescence in all channels, making the visualisation of the FISH probes difficult. Two methods were used to try and remove the autofluorescence: H_2O_2 treatment [19] and toluidine blue O [20].

H_2O_2 treatment

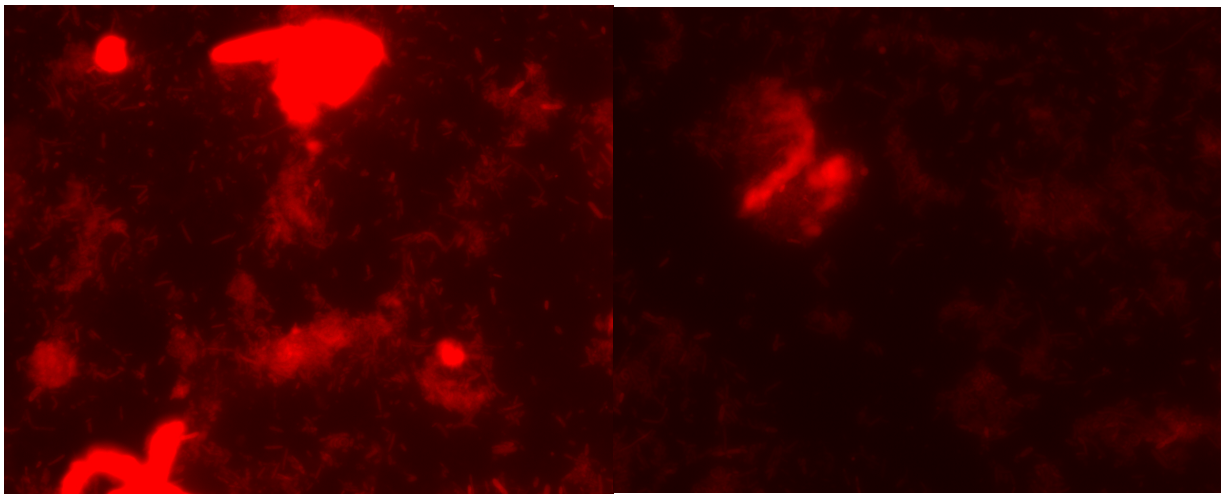
Alcoholic 6% H_2O_2 solution was prepared by combining one volume of 30% H_2O_2 and four volumes of 100% ethanol. The fixed koala cells were thoroughly washed with 100% ethanol and incubated in alcoholic H_2O_2 solution until they were decolourised (5 days). The cells were washed with 100% ethanol and subjected to microscopic observation.

Toluidine blue O

The fixed koala sample was hybridised following the FISH method above, followed by staining with 400 µl of toluidine blue O (0.05% [wt/vol]) in sterilized distilled water with 0.9M NaCl). The samples were dyed with toluidine blue O for 15 mins at room temperature and rinsed in distilled water until the water became clear. After being air dried, the samples were incubated in 99.5% ethanol for 1.5 mins to remove the dye from the bacterial cells but not the plant material. Then the samples were immediately washed with distilled water.

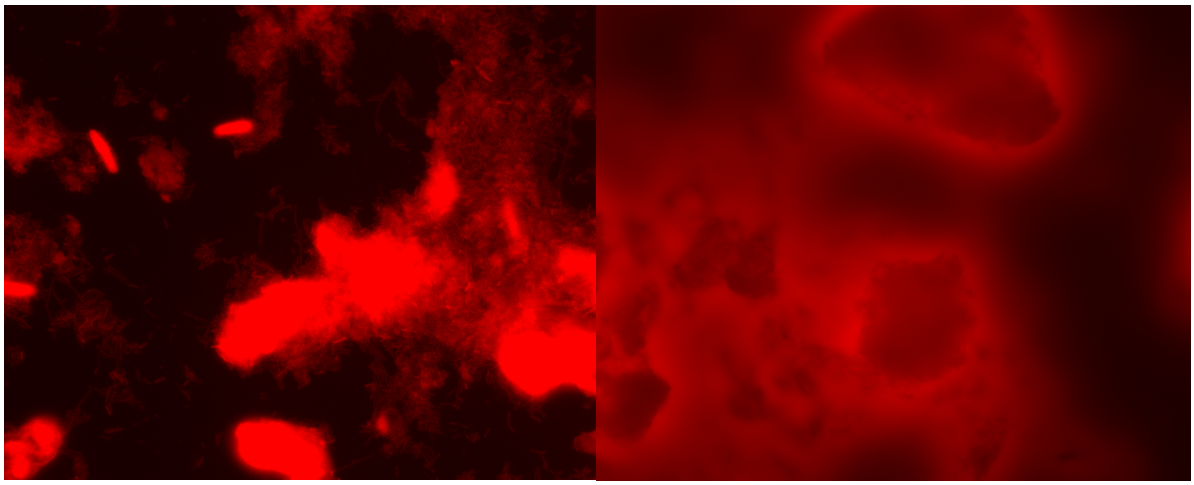
Results

Figure 9. FISH images of koala faecal samples without H_2O_2 treatment (*left*) and with H_2O_2 treatment (*right*)



FISH slides were viewed with a Nikon H600L microscope in the Cy3 channel. Both samples were taken at 60x magnification.

Figure 10. FISH images of koala faecal samples without toluidine blue O treatment (*left*) and with toluidine blue O treatment (*right*)



FISH slides were viewed with a Nikon H600L microscope in the Cy3 channel. The sample with treatment is taken at 40x magnification and the sample without treatment is taken at 60x magnification.

Conclusion

The two methods to remove autofluorescence did not work. It would be preferable to use samples collected from mammals in which their main diet does not consist of plant material.

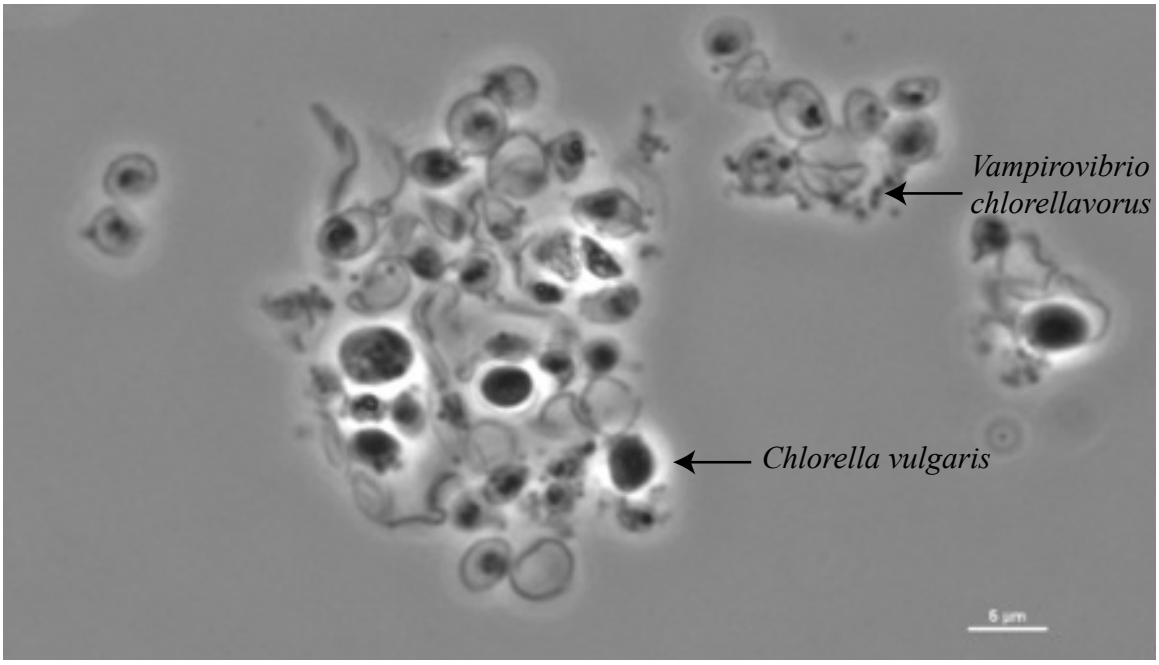
Visualising *Vampirovibrio chlorellavorus*

Methods

Vials of mixed co-cultures of *Vampirovibrio chlorellavorus* and *Chlorella vulgaris* were obtained from NIMB as NIB11383. A small amount of the lyophilised cells were re-hydrated in MQ water and placed on a microscope slide. The sample was placed under a Nikon H600L microscope under 100x magnification.

Figure 11. Vials of NIB11383 that was deposited to NIB on the 16th Feburary 1978

Figure 12. Microscope images of NIB11383



An attempt to culture the lyophilised cells was also performed as per the ATCC product sheet.

Chlorella vulgaris was grown on peptone/glucose (PG) agar (1% peptone difco, 1% glucose, 0.5% yeast extract and 1.5% Bacto agar for 7 days at room temperature.

ATCC medium:1024 MED IIa was made as below:

Tris Buffer Stock Solution (see below)	1 ml
3.33% $MgSO_4.7H_2O$ solution	100 µl
5% $CaCl_2$ solution	1 ml
Distilled water to	100 mls

Autoclave at 121°C for 20 mins

Tris buffer stock solution:	
Trizma HCl	3.5 g
Trizma base	0.335 g
Distilled water	50 mls

A loop of the *V. chlorellavorus* co-culture was resuspended in 300 µl of 1024 MedIIa and then placed in a 50 ml falcon tube with 10 mls of 1024 MedIIa and a loopful of the 7-day old *C.*

vulgaris. Cells were shaken at 100 rpm at 26°C and samples were visualised under the microscope every 2 days to identify if there was any signs of clumping of the algal cells.

Conclusion

Our attempts to culture *V. chlorellavorus* failed. This may be due to the samples not being stored correctly, as it is difficult to grow cyanobacterial cells from lyophilised material (*see Chapter 3*).

References

1. Caporaso, J.G., et al., Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J, 2012. **6**(8): p. 1621-1624.
2. Carneiro, M., et al., Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics, 2012. **13**(1): p. 375.
3. Pilloni, G., et al., Testing the Limits of 454 Pyrotag Sequencing: Reproducibility, Quantitative Assessment and Comparison to T-RFLP Fingerprinting of Aquifer Microbes. PLoS ONE, 2012. **7**(7): p. e40467.
4. Sogin, M.L., et al., Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(32): p. 12115-12120.
5. Matsuki, T., et al., Use of 16S rRNA Gene-Targeted Group-Specific Primers for Real-Time PCR Analysis of Predominant Bacteria in Human Feces. Applied and Environmental Microbiology, 2004. **70**(12): p. 7220-7228.
6. Caporaso, J.G., et al., QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 2010. **7**(5): p. 335-6.
7. Bragg, L., et al., Fast, accurate error-correction of amplicon pyrosequences using Acacia. Nature Meth, 2012. **9**(5): p. 425-426.
8. Johnson, M., et al., NCBI BLAST: a better web interface. Nucleic Acids Res, 2008. **36**: p. W5-9.
9. DeSantis, T.Z., et al., Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology, 2006. **72**(7): p. 5069-72.
10. McDonald, D., et al., An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J, 2012. **6**(3): p. 610-8.

11. Ludwig, W., et al., ARB: a software environment for sequence data. Nucleic Acids Res, 2004. **32**(4): p. 1363-71.
12. Lane, D.J., 16S/23S rRNA seqeuncing, in Nucleic acid techniques in bacterial systematics, E. Stackebrandt and M. Goodfellow, Editors. 1991, John Wiley and Sons: New York. p. 115-175.
13. Stahl, D.A., et al., Use of phylogenetically based hybridization probes for studies of ruminal microbial ecology. Applied and Environmental Microbiology, 1988. **54**(5): p. 1079-1084.
14. Heid, C.A., et al., Real time quantitative PCR. Genome Res, 1996. **6**(10): p. 986-94.
15. Dove, S.G., et al., Future reef decalcification under a business-as-usual CO(2) emission scenario. Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(38): p. 15342-15347.
16. Moter, A. and U.B. Gobel, Fluorescence in situ hybridization (FISH) for direct visualization of microorganisms. Journal of Microbiological Methods, 2000. **41**(2): p. 85-112.
17. Hugenholtz, P., Exploring prokaryotic diversity in the genomic era. Genome Biology, 2002. **3**(2): p. reviews0003.1 - reviews0003.8.
18. Yilmaz, S., et al., Fixation-free fluorescence in situ hybridization for targeted enrichment of microbial populations. ISME J, 2010. **4**(10): p. 1352-6.
19. Koga, R., T. Tsuchida, and T. Fukatsu, Quenching autofluorescence of insect tissues for in situ detection of endosymbionts. Applied Entomology and Zoology, 2009. **44**(2): p. 281-291.
20. Shinkai, T. and Y. Kobayashi, Localization of Ruminal Cellulolytic Bacteria on Plant Fibrous Materials as Determined by Fluorescence In Situ Hybridization and Real-Time PCR. Applied and Environmental Microbiology, 2007. **73**(5): p. 1646-1652.